Decoding the Regulatory Genome: Quantitative Analysis of Transcriptional Regulation in *Escherichia coli*

Thesis by
Stephanie Loos Barnes

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2018
Defended 4 May 2018

Stephanie Loos Barnes
ORCID: 0000-0002-5237-603X

ACKNOWLEDGEMENTS

During my time at Caltech I have come to believe that science is one of the humanities. Science is conducted by people for people, whether we focus on applications that directly impact quality of life or focus on the beautiful process of trying to understand the natural world and communicate our insights to others. This attitute towards science has been shaped in no small part by my advisor, Professor Rob Phillips, who has amazed me with his ability to connect with other scientists throughout the world and his deep knowledge of science history. He has remained continually supportive throughout my scientific journey.

Of course, a lab is nothing without labmates, and I've landed with some of the best. Life in the Phillips lab has been characterized by interactions between a stunning array of personalities and talents, and an advisor who is just crazy enough to put us in close quarters with one another on adventures around the world. We've been seasick in front of each other on a wildly rocking fishing boat in Alaska; we've sweated profusely in the jungles of Indonesia; we've struggled through attempts to order food and navigate cities in South Korea. All of these experiences have brought us closer together as a lab, and I believe this has made it possible to sustain the intense level of collaboration that characterizes the Phillips lab. Chapter 2 would never have been written if we hadn't formed a "science commune" of five equal authors (myself, Nathan Belliveau, Griffin Chure, Tal Einav, and Manuel Razo). Chapters 3 and 4 were built out of contributions from me, Nathan Belliveau, Bill Ireland, and our perennial collaborator from Cold Spring Harbor Laboratories, Justin Kinney. My advisor and labmates also demonstrate a continual commitment to good teaching, which has been vital in leading me towards a career as a teacher.

I am also lucky to have a wonderful network of friends and family outside of the lab. To my parents, you have supported me always and never stopped believing in my ability to do great work. To my husband Greg, you have witnessed all the ups and downs of life as a graduate student and I'm grateful that you still seem to like me anyway. You have given me a wonderful extended family who have helped make Los Angeles feel like home. Plus, you introduced me to choral singing, and without the Caltech Glee Club and the Lake Avenue Church choir I may never have discovered that vital part of myself that loves making music. To the Gloup, I've been able to see you less and less over the years, but visiting you has always felt like coming home. To Spencer, thank you for letting me eat your Cheetos.

ABSTRACT

Over the past decades DNA sequencing has become significantly cheaper and faster, which has enabled the accumulation of a huge amount of genomic data. However, much of this genomic data is illegible to us. For noncoding regions of the genome in particular, it is difficult to determine what role is played by specific DNA sequences. Here we focus on regions of DNA that play a role in transcriptional regulation. We develop models and techniques that allow us to discover new regulatory sequences and better understand how DNA sequence determines regulatory output.

We start by considering how quantitative models serve as a powerful tool for testing our understanding of biological systems. We apply a statistical mechanical framework that incorporates the Monod-Wyman-Changeux model to analyze the effects of allostery in simple repression, using the *lac* operon as a test case. By fitting our model to experimental data, we are able to determine the values of the unknown parameter values in our model. We then show that we can use the model to accurately predict the induction responses of an array of simple repression constructs with a variety of repressor copy numbers and repressor binding energies.

Next, we consider how the DNA sequence of a promoter region can provide details about how the promoter is regulated. We begin by describing an approach for discovering regulatory architectures for promoters whose regulation has not previously been studied. We focus on six promoters from *E. coli* including three well-studied promoters (*rel*, *mar*, and *lac*) to serve as test cases. We use the massively parallel reporter assay Sort-Seq to identify transcription factor binding sites with base-pair resolution, determine the regulatory role of each binding site, and infer energy matrices for each binding site. Then, we use DNA affinity chromatography and mass spectrometry to identify each transcription factor.

We conclude with an *in vivo* approach for analyzing the sequence-dependence of transcription factor binding energies. Again using Sort-Seq, we show that we can represent transcription factor binding sites using energy matrices in absolute energy units. We then show that these energy matrices can be used to accurately predict the binding energies of mutated binding sites. We provide several examples of how understanding the relationship between DNA sequence and transcription factor binding provides us with a foundation for addressing additional scientific topics, such as the co-evolution of transcription factors and their binding sites.

PUBLISHED CONTENT AND CONTRIBUTIONS

1. Nathan M. Belliveau, Stephanie L. Barnes, William T. Ireland, Daniel L. Jones, Michael Sweredoski, Annie Moradian, Sonja Hess, Justin B. Kinney, and Rob Phillips. A systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proceedings of the National Academy of Sciences*, In press, 2018.

2. Manuel Razo-Mejia, Stephanie L. Barnes, Nathan M. Belliveau, Griffin Chure, Tal Einav, Mitchell Lewis, and Rob Phillips. Tuning transcriptional regulation through signaling: A predictive theory of allosteric regulation. *Cell Systems*, In press, 2018.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

It has been said that we live in the "genomic era," a time where we can readily sequence the human genome (or any other genome, if we can get a sufficient DNA sample) [1]. However, it remains difficult to interpret the information within a genome. This is especially true of non-coding sequences such as promoters, for which there is no straightforward regulatory "code." Our ability to interpret these non-coding sequences relies on our understanding of how DNA sequence directs activities such as transcriptional regulation. This understanding is aided significantly when we can devise quantitative models that predict what will occur if the DNA sequence is altered.

In this dissertation, I focus on the challenge of creating an effective theory-experiment dialogue for dissecting transcriptional regulation. My work is built on a foundation of previous work from the Phillips lab which aimed to address the same challenge [2–11]. As a lab we are driven by the philosophy that a system cannot be considered well-understood if one cannot make falsifiable predictions about the system using generalizable models. By engaging with the many challenges of modeling transcriptional regulation, we learn information that is useful to the study of transcription and the interpretation of DNA sequences. More broadly, we help establish a quantitative approach to biology that can be applied to any system, large or small.

In this chapter, I provide a background for the work that will be presented in Chapters 2-4. I begin by discussing the role of transcription in the central dogma of molecular biology and reviewing the various mechanisms by which cells regulate gene activity. Next, I provide a primer on techniques for modeling transcriptional regulation using statistical mechanics. This approach to modeling transcriptional regulation is used throughout this dissertation. I then discuss what is known about transcriptional regulation in *E. coli*, and make the argument that there is much left to discover even for this highly-studied model organism. Finally, I provide an overview of the experimental and computational techniques used in this work.

I proceed in Chapter 2 with a thorough analysis of a thermodynamic model for a transcriptional regulatory system that incorporates the concept of allostery. In allosteric regulation, a ligand can act as a signal that stabilizes an active or inactive

form of a transcription factor. My coauthors and I present a general theory of allosteric transcriptional regulation using the Monod-Wyman-Changeux model [12]. We then rigorously test this model over a broad range of experimental parameter space using a promoter construct in which an allosteric repressor may bind to the promoter to prevent transcription. As shown in Figure 1.1A, we fit the model to a single data set to determine the values of a pair of parameters with previously unknown values. These were the only unknown parameter values for this system that were relevant to the model. With these parameter values in hand we were able to accurately predict the induction responses of multiple bacterial strains.

The aims of the study described in Chapter 2 are twofold. First, by comparing the predictions of our theoretical model against experimental measurements, we aim to confirm whether the assumptions underlying our model accurately describe allosteric transcriptional regulation. Second, we contrast our approach of predictive modeling against the common approach of descriptive modeling, in which a model such as a Hill function is fitted to the data after the fact. While descriptive models of this sort may be useful for identifying certain elements of the system, such as cooperativity between ligand binding sites, our predictive modeling approach allows us to predict how any perturbation to system parameters alters the system's phenotypic response. This ability to predict a system's response to perturbations implies a deep understanding of the primary inputs that determine the system's output.

In order to write a predictive thermodynamic model for a promoter, it is necessary to know the promoter's regulatory structure. A promoter's regulatory structure consists of its specific arrangement of transcription factor binding sites and the interactions between the transcription factors that bind to these sites. For the majority of promoters, however, the regulatory structure is unknown or only partially known, which limits rigorous modeling to synthetic promoters or the handful of natural promoters for which we definitively know the regulatory structure. In Chapter 3 we address this problem by introducing a novel approach to discovering regulatory structures for individual promoters. We use the assay Sort-Seq [13] in conjunction with DNA affinity chromatography and mass spectrometry to find transcription factor binding sites, determine the transcription factors' identities, and ascertain whether each transcription factor behaves as an activator or a repressor. We identify and test putative regulatory structures for a set of promoters which previously had no regulatory annotation. Figure 1.1B shows an example of a regulatory architecture

that was inferred using this method for a previously-unannotated regulatory region.

Regulatory architectures are necessary in order to make accurate predictions of transcriptional activity, but it is also necessary to know the binding energies of the transcription factors within the architecture. In Chapter 4 we discuss an approach to modeling DNA sequence-specific transcription factor binding energies *in vivo*, again using Sort-Seq. We develop models that allow us to predict the binding energy between a transcription factor and a mutated version of its binding site, using *lac* repressor as a test case. Figure 1.1C shows the agreement between predictions made using these models and experimental measurements of binding energies for an array of binding site mutants. We then show that this modeling technique can be used to address a number of scientific questions. For example, we observe how transcription factor sequence specificity changes when amino acid mutations are made to the transcription factor's DNA binding domain, which helps us to understand how transcription factors and their binding sites co-evolve. This provides yet another example of the importance of quantitative models for deeply understanding biological mechanisms.

In total, this work presents significant progress toward the goal of being able to dissect and quantitatively model the regulation of any promoter at will. My coauthors and I show that our predictive models provide insights beyond the outputs of specific regulatory constructs. We demonstrate our ability to identify new regulatory architectures and analyze the energetics of transcription factor binding *in vivo*. Additionally, I hope that this work will serve as a foundation for future studies that increase the scope and throughput of regulatory architecture analysis.

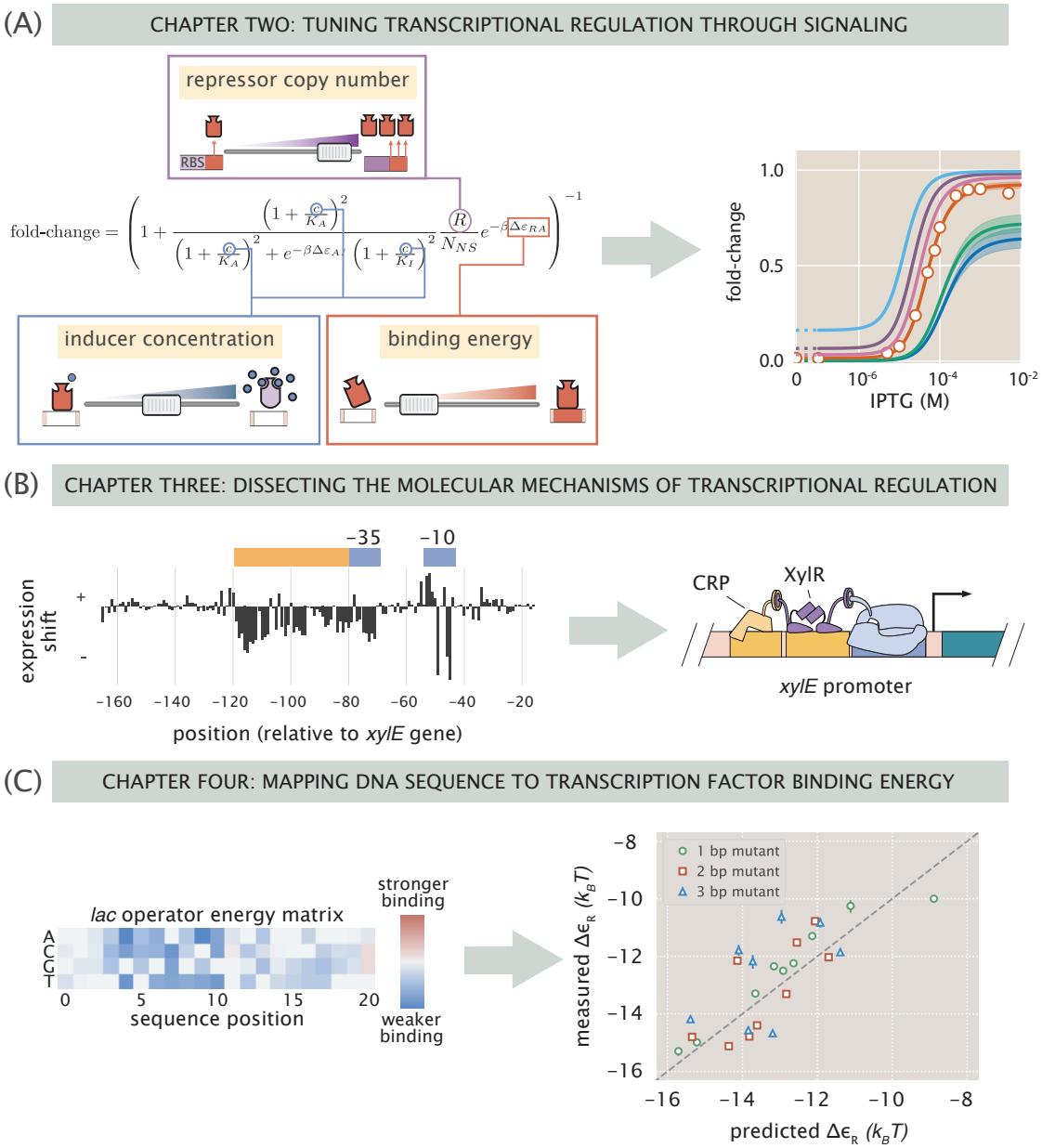

Figure 1.1: **Quantitative modeling of transcriptional regulation.** (A) In Chapter 2 we devised a model for simple repression with induction (left), which we used to predict induction responses for strains with a variety of regulatory parameters. (B) In Chapter 3 we identified regulatory architectures for unannotated promoters. We quantified the shift in expression due to mutations at each sequence position in the promoter (left) and combined these observations with DNA affinity chromatography and mass spectrometry to infer regulatory architectures (right). (C) In Chapter 4 we used *in vivo* techniques to infer energy matrices in absolute energy units, specifically k_BT (left). We used these energy matrices to predict the binding energies of binding site mutants and confirmed our predictions with experimental measurements (right).

1.1 The central dogma of molecular biology

As first described by Francis Crick [14], the central dogma of molecular biology describes how information is transferred from DNA to proteins that carry out the essential tasks of running and maintaining a cell. Figure 1.2 illustrates the basic elements of the central dogma.

The transfer of genetic data begins in the genome itself, where genes reside as regions of double-stranded DNA that code for proteins. It is important to note that not all DNA in the genome acts as a gene. A gene must be organized into a series of three-nucleotide groups (codons) that can be translated into amino acids, beginning with a "start" codon that signals the beginning of the gene and ending with a "stop" codon that signals the end of the gene. In order for a gene to be functional, it also requires sequence elements that allow it to be recognized by the protein complexes that read genetic data (RNA polymerases) and build proteins (ribosomes), as described below.

To translate a gene's DNA code into a protein, it must first be copied into a message that can be read by the ribosomes that build proteins. In a process known as transcription, an RNA polymerase (RNAP) recognizes and binds to a region upstream of the gene known as a promoter, and then copies the gene into a single-stranded RNA message known as mRNA.

Next, the mRNA is read by a ribosome. The ribosome facilitates the matching of the mRNA message to transfer RNAs (tRNAs), which are structures made out of RNA that include a codon recognition sequence on one end and carry the corresponding amino acid as cargo on the other end. The ribosome moves along the mRNA codon by codon. When it encounters a tRNA that correctly matches the current codon, it removes the amino acid from the tRNA and adds it to a growing polypeptide chain. Once the polypeptide chain is complete, it self-assembles into three-dimensional protein structure.

The remarkable thing about this process is that it is nearly identical for all organisms on earth, hence earning it the title of "central dogma." There are, however, a few differences between prokaryotes and eukaryotes that are worth noting. First, eukaryotic genes include both coding and non-coding regions. The non-coding regions must be removed after translation so that the ribosome can read a single coherent mRNA message. Prokaryotic genes are much more concise by contrast, as they only include coding DNA. Another notable difference is in the organization of the genome itself. Eukaryotic genomes are typically distributed across a number of chromosomes in

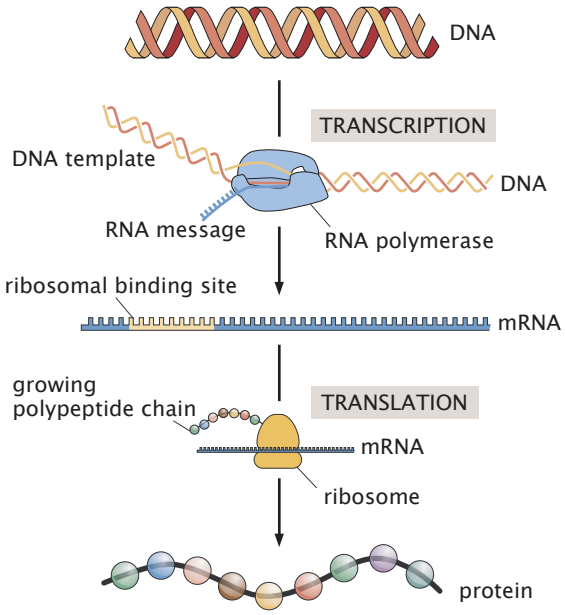

Figure 1.2: **The central dogma of molecular biology.** Genes are encoded as DNA sequences within the genome. RNA polymerase (RNAP) copies the DNA as a single-stranded mRNA transcript. Then, ribosomes translate the mRNA into protein by facilitating the pairing of tRNAs with the mRNA transcript and joining the associated amino acids together into a polypeptide chain. This polypeptide chain then generally self-assembles into a protein.

which the DNA is tightly wound around clusters of protein complexes known as histones to create a DNA packaging complex known as a nucleosome. Importantly, DNA must be unwound from a nucleosome in order to be accessible for transcription. Prokaryotes have smaller genomes that are often contained on a single circular chromosome, and while structural proteins do play some role in organizing the genome, the DNA is not as tightly sequestered as it is in eukaryotic nucleosomes and is generally available for transcription. In spite of these differences, the central dogma accurately describes the transfer of information from DNA to protein for both prokaryotes and eukaryotes. In this work we primarily focus on prokaryotes, and we note that while eukaryotic genomes possess some complicating factors beyond what is encountered in prokaryotic genomes, the essential principles governing the prokaryotic genome can be applied almost without reservation to the eukaryotic genome.

1.2 Regulation of protein abundance and activity

As cells enter different environmental conditions or growth states, the copy numbers and activities of proteins are altered to suit the cells' changing needs. This is true for both prokaryotes and eukaryotes [15–19]. It applies to global changes such as increases in ribosome copy number as a cell's growth rate increases [16] as well as highly specific changes like the activation of the *lac* operon in the presence of allolactose [20]. There are many mechanisms for controlling protein copy number and activity, and control can be enacted at any stage of the central dogma: during transcription, post-transcription, or even after proteins have been synthesized. Figure 1.3 illustrates examples of regulation at each stage of the central dogma.

Transcriptional regulation refers to any regulatory mechanism that either helps or hinders the process of creating an mRNA transcript. In general this is accomplished by modulating the probabilty that RNAP will bind to the promoter and proceed to copy the gene. The probability of RNAP binding depends in part on the sequence of the promoter itself, as the polymerase and any associated sigma factors have DNA sequence binding preferences and deviating from these preferences will reduce the probability of binding. However, the promoter sequence is static and cannot respond to changes in environment or growth state. In order to enact transcriptional regulation that can change in response to a stimulus, the cell produces DNA-binding proteins known as transcription factors that bind to the promoter near the RNAP binding site and modulate RNAP binding activity. For example, in a regulatory scheme known as simple activation, a transcription factor known as an activator may bind to the DNA immediately upstream of the RNAP. The activator interacts with RNAP in a manner that encourages RNAP binding, effectively decreasing the free energy of RNAP binding and raising the probability that RNAP will bind and initiate transcription. Similarly, in simple repression a transcription factor known as a repressor binds within or near the RNAP binding site so that it blocks the RNAP from binding, thereby reducing the probability of RNAP binding.

Post-transcriptional regulation refers to any modification made to the mRNA transcript that affects the number of proteins that can be copied from the transcript. Post-transcriptional regulation can differ significantly between eukaryotes and prokaryotes. In eukaryotes, the mRNA must be prepared for translation by removing non-coding segments and exporting the mRNA from the nucleus. Any alterations to the mRNA that interfere with these processes can regulate protein abundance by preventing translation. Because prokaryotes do not have these processes, many mecha-

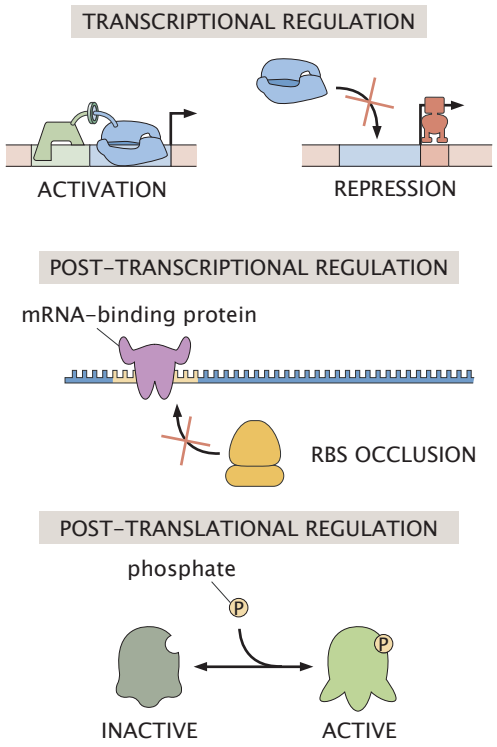

Figure 1.3: **Regulation of protein abundance and activity at each stage of the central dogma.** Gene expression and activity can be regulated at any stage in the process of translating a gene into a protein. For example, transcription can be regulated when a transcription factor such as an activator binds to the DNA near the RNA polymerase binding site and modulates the probability of RNAP binding. A gene can also be regulated post-transcription, for instance when an mRNA-binding protein binds to the ribosomal binding site and prevents a ribosome from binding and translating the gene. Even after proteins have been translated, their activity can be modulated by modifications such as phosphorylation.

nisms for post-transcriptional regulation in eukaryotes do not occur in prokaryotes. However, some forms of post-transcriptional regulation are present in prokaryotes. For example, an mRNA-binding protein such as CsrA may bind at or near the ribosomal binding site and prevent the ribosome from binding to and translating the mRNA, similar to simple repression in transcriptional regulation [21]. Another common mechanism for post-transcriptional regulation in prokaryotes is the use of small RNA regulators (sRNAs). These RNAs possess extensive complementarity to their target mRNAs and modulate their activity by either binding to the mRNA and sequestering it (often with the help of a chaperone protein such as Hfq) or by directing cleavage of mRNAs by RNAse [22].

Finally, post-translational regulation is any modification to a protein that modulates

its activity. One very common example of this is phosphorylation, whereby the covalent attachment of a phosphate to a protein instigates a change in the protein's structure that switches it between an "active" and "inactive" state. Phosphorylation often occurs in response to some external signal. For example, in two-component signal transduction, a small molecule serves as a signal indicating some environmental change. The small molecule binds noncovalently to the receptor region of a membrane protein and induces the transfer of a phosphate from the membrane protein to a second protein that creates some response [23]. Alternatively, a small molecule can bind directly to an allosteric protein, thus inducing a structural change and creating a response to the signal. Allostery is a key component of the present work and is discussed in more detail in the next section.

We mention examples of regulation at each stage of the central dogma to give a sense of the complex interplay of stimuli and responses that determine a cell's state at any given time. While mechanisms of regulation at all stages of the central dogma certainly inform one another, the focus of this work is the quantitative analysis of transcriptional regulation in prokaryotes. In the following sections we will go into greater detail regarding the mechanisms of regulation during transcription.

1.3 Quantitative models of transcriptional regulation

A core principle of this work is the power of quantitative modeling for developing an understanding of the natural world. Quantitative models allow us to make predictions regarding a system's behavior. Additionally, analytically-derived quantitative models allow us to test our understanding of the essential mechanisms that drive a system.

As an example of how models can be used to test assumptions, for transcription we typically use models that rely on the assumption that the probability of RNAP binding to the promoter, p_{bound}, is proportional to gene expression. This assumption relies on the "occupancy hypothesis" as utilized in Ref. [24]. The occupancy hypothesis posits that binding of RNAP or a transcription factor to a binding site indicates that the protein is actively playing a role in transcription. This means that for RNAP, occupancy of a promoter implies that transcription is taking place; for a transcription factor, occupancy of a binding site implies that transcriptional regulation is taking place. We often apply this assumption when writing models for transcriptional regulation (see model in Chapter 2), but it is not always valid. In Ref. [6] it was found that the occupancy hypothesis could not adequately describe the mechanism of repression in a particular regulatory architecture. In this architecture, RNAP and a repressor could bind simultaneously such that the repressor prevented RNAP from proceeding with transcription. Occupancy of RNAP at this promoter did not necessarily imply that transcription would take place. This violation of the occupancy hypothesis was identified by comparing gene expression data to predictions from a model that used the occupancy hypothesis, and observing that the model did not adequately describe the data.

In this section we will demonstrate how models can be used to represent constitutive transcription, and then show how one can generalize these models to include additional regulatory mechanisms. We provide full derivations for several models. Note that these derivations are discussed in detail elsewhere [2, 25], but are reproduced here for the benefit of the reader.

A statistical mechanical approach to modeling constitutive transcription

Here we will consider a statistical mechanical approach for modeling constitutive transcription. Statistical mechanics concerns itself with the probability of different microstates in systems containing a large number of interacting particles. A microstate is a unique arrangement of particles, which may or may not have prop-

erties that are distinguishable from other microstates. The probability of a specific microstate is given by the Boltzmann distribution,

$$p(\varepsilon_i) = \frac{1}{Z} e^{-\beta \varepsilon_i}, \tag{1.1}$$

where ε_i is the energy of microstate i, Z is the partition function (as described below), and β is equal to $1/k_B T$ where k_B is Boltzmann's constant and T is the temperature of the system. The quantity $e^{-\beta \varepsilon_i}$ is referred to as the Boltzmann factor. The partition function can be thought of the sum of the statistical mechanical weights of all microstates in the system, and is given by

$$Z = \sum_{i=1}^{N} e^{-\beta \varepsilon_i}. \tag{1.2}$$

When modeling transcription, our goal is to determine the probability that an RNAP will bind to a promoter and initiate transcription. When using a statistical mechanical approach, we identify the various states that a system can adopt, where a state is a set of microstates with indistinguishable properties. We assign statistical mechanical weights to each state and use these weights to determine the probability of RNAP binding, p_{bound}. This identification of states and weights is modeled for the case of constitutive transcription in Figure 1.4. Here we provide a derivation of a statistical mechanical expression for the probability of RNAP binding at a constitutive promoter. We show how this derivation can be represented by a states and weights diagram, which can then be used to greatly simplify the process of deriving models for more complex regulatory scenarios.

In the case of constitutive transcription, the system's interacting particles are the cell's many copies of RNAP and the many DNA binding sites available to the RNAP. A microstate can be thought of as a "snapshot" of the positions of all RNAP relative to the genome at a given time. If we are interested in the transcription of a specific gene, then we wish to know the probability that a single copy of RNAP is bound to that gene's promoter. We can determine this probability using Equation 1.1 provided we know (A) the energy ε_i of the state, (B) the multiplicity of the state (i.e., the number of possible microstates in which RNAP is bound to the promoter of interest) and (C) the partition function Z that represents all possible microstates of the system.

To simplify the problem, we abstract the genome as a single specific RNAP binding site and a series of nonspecific binding sites that bind weakly with the RNAP. In reality, there are many specific RNAP binding sites in the genome, and any given stretch of DNA will have a unique RNAP binding energy that ranges from very weak to very strong. For the purpose of this problem, however, we can view all DNA aside from our binding site of interest as being part of a "pool" of DNA binding sites with some average weak binding energy. There are N_{NS} of these nonspecific binding sites, where we assume that N_{NS} is approximately equal to the length of the genome. We assign an energy of ε_P^S to an RNAP bound to the specific binding site and ε_P^{NS} to an RNAP bound to any of the nonspecific sites.

The energy of any microstate i in which an RNAP is bound to the specific site must account for both the energy of one RNAP binding to the specific site and $P-1$ RNAPs binding to nonspecific sites, where P is the total number of RNAPs in the system, such that $\varepsilon_i = (P-1)\varepsilon_P^{NS} + \varepsilon_P^S$. The Boltzmann factor for such a microstate is thus $e^{-\beta(P-1)\varepsilon_P^{NS}} e^{-\beta\varepsilon_P^S}$. The value of p_{bound} is given by the sum of the Boltzmann weights for all microstates in which an RNAP is bound to the specific site, giving us

$$p_{bound} = \frac{\sum_{i=1}^{N} e^{-\beta(P-1)\varepsilon_P^{NS}} e^{-\beta\varepsilon_P^S}}{Z_{tot}}, \tag{1.3}$$

where we define Z_{tot} as the total partition function of the system. We can re-organize p_{bound} as

$$p_{bound} = \frac{e^{-\beta\varepsilon_P^S} \sum_{i=1}^{N} e^{-\beta(P-1)\varepsilon_P^{NS}}}{Z_{tot}}, \tag{1.4}$$

which can then be rewritten as

$$p_{bound} = \frac{e^{-\beta\varepsilon_P^S} Z_{NS}(P-1, N_{NS})}{Z_{tot}}, \tag{1.5}$$

where $Z_{NS}(P-1, N_{NS})$ is a partial partition function representing all microstates in which $(P-1)$ RNAP are distributed among N_{NS} nonspecific binding sites. We can further recognize that Z_{tot} is the sum of all microstates in which an RNAP is bound

to the specific site and all microstates in which no RNAP is bound to the specific site, such that $Z_{tot} = e^{-\beta\varepsilon_P^S} Z_{NS}(P-1, N_{NS}) + Z_{NS}(P, N_{NS})$, which gives us

$$p_{bound} = \frac{e^{-\beta\varepsilon_P^S} Z_{NS}(P-1, N_{NS})}{e^{-\beta\varepsilon_P^S} Z_{NS}(P-1, N_{NS}) + Z_{NS}(P, N_{NS})}. \tag{1.6}$$

We can now see that the Equation for p_{bound} is composed of a set of partition functions in which each partition function represents the statistical mechanical weight of one of the system's states. In the case of constitutive transcription, the states are either *bound*, which consists of all microstates in which an RNAP is bound to the specific site and has a weight given by $e^{-\beta\varepsilon_P^S} Z_{NS}(P-1, N_{NS})$, or *unbound*, which consists of all microstates in which no RNAP is bound to the specific site and has a weight given by $Z_{NS}(P, N_{NS})$. These states are represented pictorially in the "STATES" column of Figure 1.4.

Next we wish to rewrite Equation 1.6 using measurable parameters. A partition function can be thought of as the product of a state's Boltzmann factor and the state's multiplicity, or the number of microstates that comprise the state. We have already determined the Boltzmann factors for each state in our model, and the multiplicities can be determined combinatorially. This gives us the statistical mechanical weight of the bound state,

$$e^{-\beta\varepsilon_P^S} Z_{NS}(P-1, N_{NS}) = \frac{(N_{NS})!}{(P-1)!(N_{NS}-P+1)!} e^{-\beta(P-1)\varepsilon_P^{NS}} e^{-\beta\varepsilon_P^S}, \tag{1.7}$$

and the statistical mechanical weight of the unbound state,

$$Z_{NS}(P, N_{NS}) = \frac{(N_{NS})!}{P!(N_{NS}-P)!} e^{-\beta P\varepsilon_P^{NS}}. \tag{1.8}$$

These weights can be simplified using the approximation $\frac{(N_{NS})!}{P!(N_{NS}-P)!} \approx \frac{(N_{NS})^P}{P!}$ where $N_{NS} \gg P$. The simplified weights are represented in the "WEIGHTS" column of Figure 1.4. We can now rewrite p_{bound} as

$$p_{bound} = \frac{\frac{(N_{NS})^{(P-1)}}{(P-1)!} e^{-\beta(P-1)\varepsilon_P^{NS}} e^{-\beta\varepsilon_P^S}}{\frac{(N_{NS})^{(P-1)}}{(P-1)!} e^{-\beta(P-1)\varepsilon_P^{NS}} e^{-\beta\varepsilon_P^S} + \frac{(N_{NS})^P}{P!} e^{-\beta P\varepsilon_P^{NS}}}. \tag{1.9}$$

Finally, we can greatly simplify the form of the equation by dividing the weight for each state by the weight for the unbound state. The unbound state then has

a renormalized weight equal to 1, and the bound state has a renormalized weight of $\frac{P}{N_{NS}} e^{-\beta(\varepsilon_P^S - \varepsilon_P^N S)}$. We can then define $\Delta\varepsilon_P = \varepsilon_P^S - \varepsilon_P^{NS}$ where $\Delta\varepsilon_P$ represents the difference in RNAP binding energy between the specific binding site and the nonspecific genomic background. The renormalized weights for each state are illustrated in Figure 1.4 column "RENORMALIZED WEIGHTS." Substituting the renormalized values into equation 1.9 gives us our final equation for the probability of RNAP binding to a constitutive promoter,

$$p_{bound} = \frac{\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}{1 + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}. \tag{1.10}$$

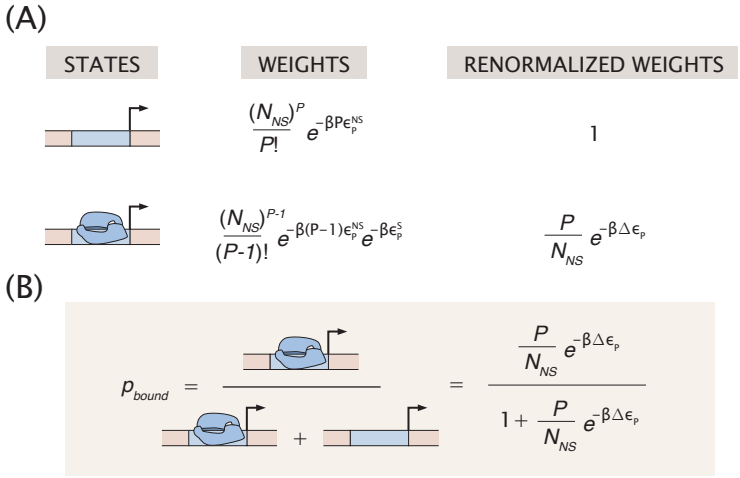

Figure 1.4: **Modeling transcription using statistical mechanics.** To model gene expression, we make the assumption that gene expression is proportional to the probability that RNAP is bound to the promoter, p_{bound} [24]. (A) To determine the value of p_{bound} we then enumerate all of the states available to the system and assign statistical mechanical weights based on the energy associated with each state and the multiplicity of each state. Renormalizing the weights such that the unbound state has a weight of 1 then provides us with a clean set of statistical mechanical weights that can be used to determine the value of p_{bound}. (B) The value of p_{bound} is equal to the statistical mechanical weight of the RNAP bound state divided by the sum of the weights of all possible states.

This equation for p_{bound} provides us with a foundation which we can build upon to create models for more complex regulatory scenarios. To provide examples of how this works, we now consider the cases of simple activation and simple repression.

Using the "states and weights" approach to model transcriptional regulation

The renormalized states and weights in Figure 1.4 reveal a pattern that can be used to easily determine states and weights for more complex architectures. Specifically, the multiplicity associated with some DNA-binding protein X can be represented as $\frac{X}{N_{NS}}$, and the Boltzmann factor associated with it can be represented as $e^{-\beta\Delta\varepsilon_x}$. Any other energies associated with any of the states, such as an interaction energy between proteins, can likewise be represented using a Boltzmann factor.

Simple Repression

We consider the case of simple repression, in which a repressor binds adjacent to an RNAP binding site and prevents RNAP from binding. In this case there are three states available to the system: no proteins bound, repressor bound, and RNAP bound. These states and their associated weights are enumerated in Figure 1.5A.

The expression for the probability of RNAP binding in a simple repression architecture is again found by dividing the statistical weight of the RNAP bound state by the sum of the statistical weights of all states, which gives us

$$p_{bound} = \frac{\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}{1 + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P} + \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_R}}. \tag{1.11}$$

As noted previously, it is assumed that gene expression is proportional to p_{bound}. However, it is difficult to determine the exact proportionality between these quantities, and we lack a straightforward way to measure p_{bound} *in vivo*. It is therefore more convenient to model regulatory systems using the fold-change, which quantifies the change in expression due to regulation. This quantity is straightforward to measure experimentally and has a clear interpretation in regards to regulatory strength. For repression the fold-change is given by

$$\text{fold-change} = \frac{p_{bound}(R)}{p_{bound}(R = 0)}. \tag{1.12}$$

To obtain a more detailed expression for fold-change, we substitute Equation 1.11 into Equation 1.12, which gives us

$$\text{fold-change} = \left(\frac{\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}{1 + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P} + \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_R}} \right) \left(\frac{1 + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}{\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}} \right). \tag{1.13}$$

To simplify this expression, we make use of the weak promoter approximation, where we assume RNAP binds weakly to the promoter which implies that $\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P} \ll 1$. This allows us to write

$$\text{fold-change} \approx \left(\frac{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}}{1 + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_R}} \right) \left(\frac{1}{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}} \right), \tag{1.14}$$

which leads to our final expression for the fold-change of a simple repression system,

$$\text{fold-change} \approx \frac{1}{1 + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_R}}. \tag{1.15}$$

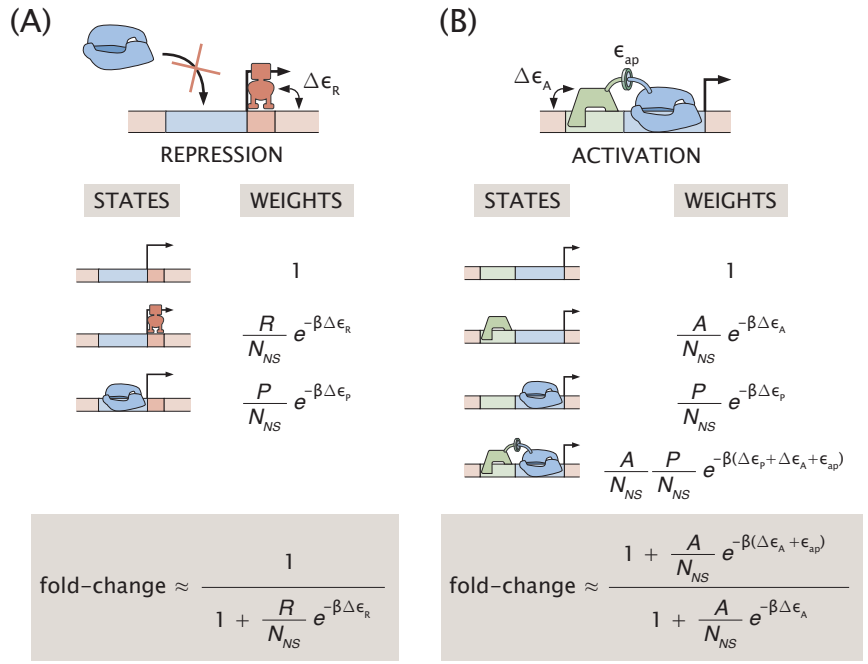

Figure 1.5: **The probability of RNAP binding is regulated by transcription factors.** Transcription factors bind to DNA within a promoter and alter the probability that RNAP will bind to the promoter and initiate transcription. (A) *Simple repression* occurs when a single transcription factor binds in the vicinity of the RNAP binding site and prevents RNAP binding. (B) *Simple activation* occurs when a single transcription factor binds in the vicinity of the RNAP binding site and promotes RNAP binding.

Simple Activation

The case of simple activation is similar to simple repression, though it incorporates the complicating factor of cooperative interactions between proteins. In simple

activation, an activator and RNAP can bind to the promoter simultaneously, as noted in the states and weights diagram for simple activation shown in Figure 1.5B. The binding of multiple proteins gives this state a multiplicity of $\frac{A}{N_{NS}}\frac{P}{N_{NS}}$, where A is the number of activators in the system. An interaction energy between the activator and RNAP, ε_{ap}, serves to solidify RNAP binding and must be included in the Boltzmann factor, which is then represented as $e^{-\beta(\Delta\varepsilon_A+\Delta\varepsilon_P+\varepsilon_{ap})}$, where $\Delta\varepsilon_A$ represents the binding energy of the activator to its binding site.

For a simple activation system, p_{bound} is given by

$$p_{bound} = \frac{\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} + \frac{A}{N_{NS}}\frac{P}{N_{NS}}e^{-\beta(\Delta\varepsilon_A+\Delta\varepsilon_P+\varepsilon_{ap})}}{1 + \frac{A}{N_{NS}}e^{-\beta\Delta\varepsilon_A} + \frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} + \frac{A}{N_{NS}}\frac{P}{N_{NS}}e^{-\beta(\Delta\varepsilon_A+\Delta\varepsilon_P+\varepsilon_{ap})}}. \tag{1.16}$$

The fold-change for simple activation is given by

$$\text{fold-change} = \frac{p_{bound}(A)}{p_{bound}(A=0)}. \tag{1.17}$$

We note that while the form of the fold-change for simple activation mirrors the form for simple repression, we expect the fold-change values for simple activation to be greater than 1 and we expect the fold-change values for simple repression to be less than 1. This is a result of the greater expression values that occur when $A > 0$ and the lower expression values that occur when $R > 0$.

To obtain a more detailed expression for the fold-change of a simple activation system, we first simplify Equation 1.16 by rewriting it as

$$p_{bound} = \frac{\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} + \frac{A}{N_{NS}}\frac{P}{N_{NS}}e^{-\beta(\Delta\varepsilon_A+\Delta\varepsilon_P+\varepsilon_{ap})}}{1 + \frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} + \frac{A}{N_{NS}}e^{-\beta\Delta\varepsilon_A}\left(1 + \frac{P}{N_{NS}}e^{-\beta(\Delta\varepsilon_P+\varepsilon_{ap})}\right)}, \tag{1.18}$$

which enables us to use the weak promoter approximations $\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} \ll 1$ and $\frac{P}{N_{NS}}e^{-\beta(\Delta\varepsilon_P+\varepsilon_{ap})} \ll 1$ to obtain

$$p_{bound} \approx \frac{\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} + \frac{A}{N_{NS}}\frac{P}{N_{NS}}e^{-\beta(\Delta\varepsilon_A+\Delta\varepsilon_P+\varepsilon_{ap})}}{1 + \frac{A}{N_{NS}}e^{-\beta\Delta\varepsilon_A}}. \tag{1.19}$$

Now we plug this expression into Equation 1.17 to get

$$\text{fold-change} \approx \left(\frac{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P} + \frac{A}{N_{NS}} \frac{P}{N_{NS}} e^{-\beta(\Delta\varepsilon_A + \Delta\varepsilon_P + \varepsilon_{ap})}}{1 + \frac{A}{N_{NS}} e^{-\beta \Delta\varepsilon_A}} \right) \left(\frac{1}{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}} \right). \quad (1.20)$$

This simplifies to the final form of the fold-change equation for simple activation,

$$\text{fold-change} \approx \frac{1 + \frac{A}{N_{NS}} e^{-\beta(\Delta\varepsilon_A + \varepsilon_{ap})}}{1 + \frac{A}{N_{NS}} e^{-\beta \Delta\varepsilon_A}}. \quad (1.21)$$

The examples of simple repression and simple activation show how statistical mechanical models can be applied to simple architectures. One can use this same approach to derive models for more complex architectures, provided that the interactions between the various elements of the architectures (e.g., transcription factors, binding sites, etc.) are sufficiently understood. Chapter 2 addresses how the model for simple repression can be adapted to account for the addition of an inducer ligand. Refs. [7, 8] apply the states and weights approach to the case of DNA looping in the *lac* operon. Ref. [2] explains further how statistical mechanical models can be applied to a broad variety of regulatory architectures.

1.4 The diversity of transcriptional regulatory mechanisms

There is great diversity in bacterial regulatory architectures, often representing sophisticated control systems that respond sensitively to environmental changes [23, 26–28]. At a basic level, we can think of regulatory architectures as arrangements of transcription factors that interact with one another and with RNAP. Figure 1.6 shows the diversity of known regulatory architectures in *E. coli* as recorded in RegulonDB. Many of the promoters in RegulonDB have no regulatory annotations, which indicates either that the promoters are constitutive or their architectures have not yet been identified. Among promoters with regulatory annotations, we see that it is most common for a promoter to have a single binding site annotation, but there remain many promoters with two or three recorded transcription factor binding sites. The data for promoters with four or more binding sites are not included in this plot.

The distribution of transcription factor binding sites can give us a sense of how a promoter is regulated, but there are a number core regulatory mechanisms that cannot be captured by mapping binding sites alone. In Figure 1.7 we use the classic example of the *lacZYA* operon to illustrate three important regulatory mechanisms that go beyond simple transcription factor binding. These mechanisms are allostery, looping, and binding by architectural proteins.

Allostery

Allostery is an exceedingly common phenomenon in both eukaryotes and prokaryotes, in multiple classes of proteins. As discussed in detail in Chapter 2, an allosteric protein switches between two or more conformations which can have different properties. Binding of a ligand to the protein can dramatically increase the probability that the protein will adopt a particular conformation. For transcription factors, this means that allostery serves as a form of "one-component signaling," where a small molecule signal directly stimulates a response such as a change in gene regulation [23].

Both CRP and LacI are allosteric transcription factors. In the case of the global regulator CRP, the small molecule cAMP must be present in order for CRP to adopt a conformation that binds to DNA with high affinity. Conversely, when the small molecule allolactose (a derivative of lactose) binds to LacI, LacI adopts a conformation that has a weak affinity for DNA. Because LacI acts as a repressor for the *lacZYA* operon, this means that allolactose acts as a signal that induces transcription of the *lacZYA* operon. This initiates production of the proteins that

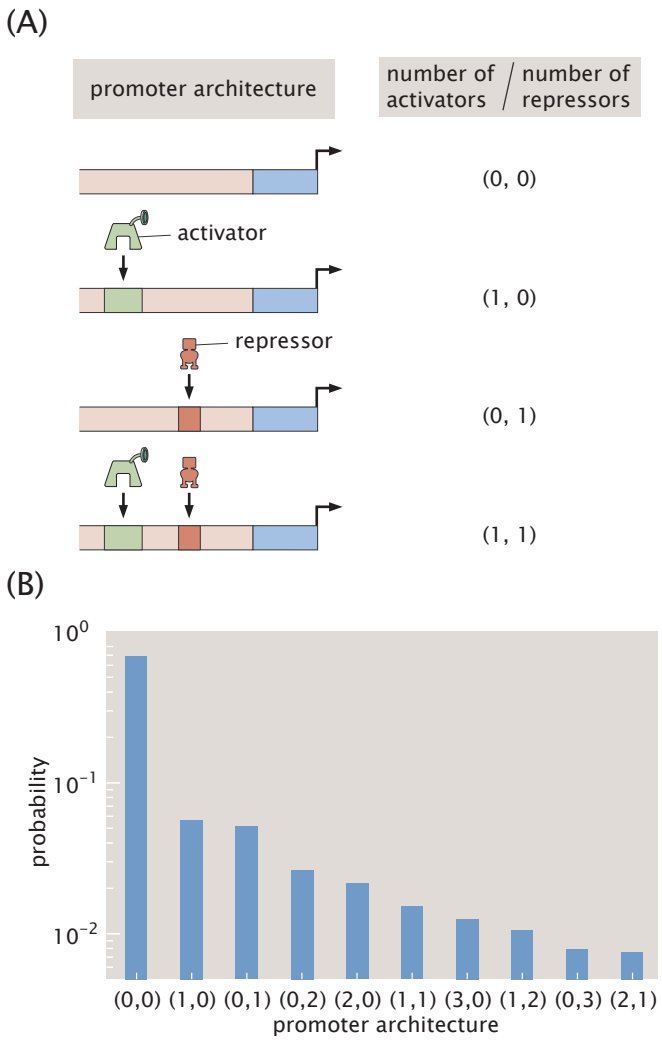

Figure 1.6: **Distribution of regulatory architectures in *E. coli*.** (A) We classify regulatory architectures according to the number of activator sites A and repressor sites R in a promoter region, using the notation (A, R). This classification does not specify the positions of the binding sites. (B) We plot the frequencies of different regulatory architectures as noted in RegulonDB. Note that many promoters lack complete regulatory annotations, which skews the data towards (0,0).

enable the cell to metabolize lactose. In laboratory settings IPTG (isopropyl β-D-1-thiogalactopyranoside) is frequently used as an inducer for the *lacZYA* operon instead of allolactose. IPTG is used because unlike allolactose, IPTG is not degraded by β-galactosidase, meaning that the concentration remains constant for the duration of the experiment.

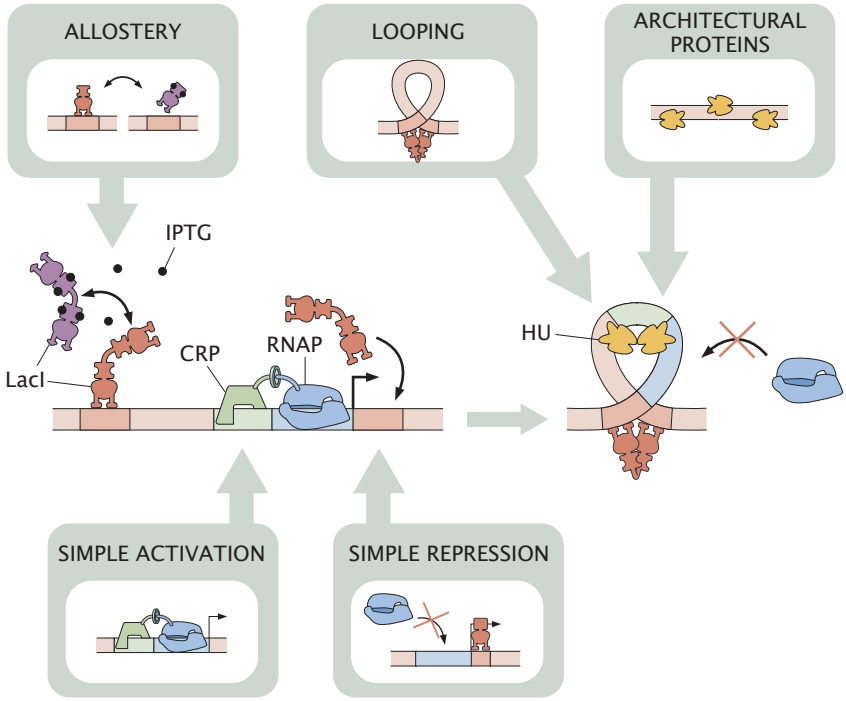

Figure 1.7: **Architecture of the *lacZYA* operon.** The *lacZYA* operon is regulated by CRP, which acts as a simple activator, and LacI, which acts as a repressor. LacI is an allosteric repressor that can adopt either an "active" conformation (red) that binds strongly to the DNA and prevents RNAP binding, or an "inactive" conformation (purple) that binds weakly to the DNA. When the ligand allolactose (or, alternatively, IPTG) binds to LacI, it stabilizes the inactive conformation and prevents repression. LacI can perform either simple repression or repression by looping. Looping is facilitated by binding of the architectural protein HU.

Looping

Looping is a form of action at a distance in which a transcription factor binds simultaneously to two binding sites that are separated by hundreds of base pairs or more, which requires the intervening DNA to form a loop. Action at a distance is a common strategy employed by eukaryotic enhancers [29]. While only a handful of looping architectures have been studied in prokaryotes, a scan of RegulonDB indicates at least $\sim$ 50 instances of binding sites for the same transcription factor that are spaced approximately 90 bp apart, which is the minimum distance observed in well-studied natural looping architectures [30]. The prevalence of looping in *E. coli* may be much higher than this, as Ref. [30] explored a narrow range of loop sizes and there are likely to be many transcription factor binding sites in the *E. coli* genome that are not currently reported in RegulonDB.

The *lacZYA* operon is a classic example of looping as a component of a regulatory architecture which has been studied extensively to better understand the physics of DNA bending in the context of gene regulation [7, 8, 31–39]. Looping in *lacZYA* promotes repression by increasing the local concentration of repressor [29] and providing an additional state in which RNAP is prevented from binding to the promoter. Repression appears to be the most common usage of looping architectures in prokaryotes, though there are examples of looping being used for activation of σ^{54}-dependent transcription in a manner similar to eukaryotic enhancers, as reviewed extensively in Ref. [40].

Architectural Proteins

Bacteria possess a class of proteins that are analogous to histones in eukaryotes. These proteins are alternatively known as architectural proteins, nucleoid-associated proteins, or histone-like proteins. Like histones, they are known to play a role in gene regulation (reviewed in Ref. [41]) and chromatin organization (reviewed in Ref. [42]). Some bind to specific DNA sequences, while others appear to bind nonspecifically or bind preferentially to bent DNA.

The architectural protein HU binds to deformed DNA and bends it. In the *lacZYA* operon, HU contributes to repression by binding to the looping region and facilitating looping between two LacI binding sites. Cells lacking HU exhibit significantly lower repression levels at the *lacZYA* operon than cells possessing HU. Furthermore, while looping mechanics are known to depend on the DNA sequence of the looping region, bending due to HU does not appear to be affected by the relative "stiffness" of the DNA [8].

We discuss these examples to give a sense of the diverse modes of transcriptional regulation in prokaryotes, and we note that this is not a comprehensive list of types of transcriptional regulation. There are a number of other schemes that are worthy of discussion and have been addressed thoroughly elsewhere, including (but not limited to) regulatory role switching as in the *araBAD* operon [27], toxin-antitoxin systems (see Chapter 3 for the example of the *relBE* operon), the role of DNA shape [43], and altering DNA specificity using methylation [44].

1.5 The state of knowledge of transcriptional regulation

Although a number of regulatory architectures and mechanisms have been subject to rigorous study, we still know very little about how most genes are regulated. For example, *E. coli* is arguably the most well-studied and well-documented model organism, yet most operons lack any regulatory annotation in databases like RegulonDB and EcoCyc [45, 46]. Figure 1.8A shows a map of the *E. coli* genome, color-coded by whether each operon has any regulatory annotation in RegulonDB. As of the present work, only 33% of operons in *E. coli* are annotated as having transcription factor binding sites in their promoter regions.

If an operon lacks any regulatory annotation, it is possible that it is constitutively expressed. However, another hypothesis is that transcription factor binding sites exist for many of these unannotated operons, but have just not been discovered yet. A look at data from previous releases of RegulonDB indicates that transcription factor binding sites are being discovered at a steady rate (see the RegulonDB Summary History at `http://regulondb.ccg.unam.mx/menu/about_regulondb/regulondb_history/database_summary.jsp`). Figure 1.8B plots the relative numbers of genes and transcription factor binding sites recorded in RegulonDB over a 10 year period (note that updates to RegulonDB occur every few months to incorporate results from new literature). While the number of genes remains fairly constant, the number of transcription factor binding sites has nearly doubled over the last 10 years and appears to be continuing to increase. This lends support to the hypothesis that an operon's lack of regulatory annotation often indicates ignorance rather than constitutive expression. If this hypothesis is correct, it also is likely that many annotated promoters have incomplete information and have more complex regulatory architectures than it would appear.

Even for operons with well-annotated regulatory regions, it can be difficult to determine how the transcription factors interact with one another to regulate gene expression. As discussed in Section 1.4, simply knowing the arrangement of activators and repressors does not always accurately capture the regulatory mechanisms in play, and tells us nothing about important features like allostery. This represents a significant gap in our understanding of transcriptional regulation. While it remains difficult to determine these details on a high-throughput scale, this work makes significant progress towards this goal by deciphering regulatory architectures, modeling complex regulatory mechanisms, and analyzing the energetics of transcription factor binding.

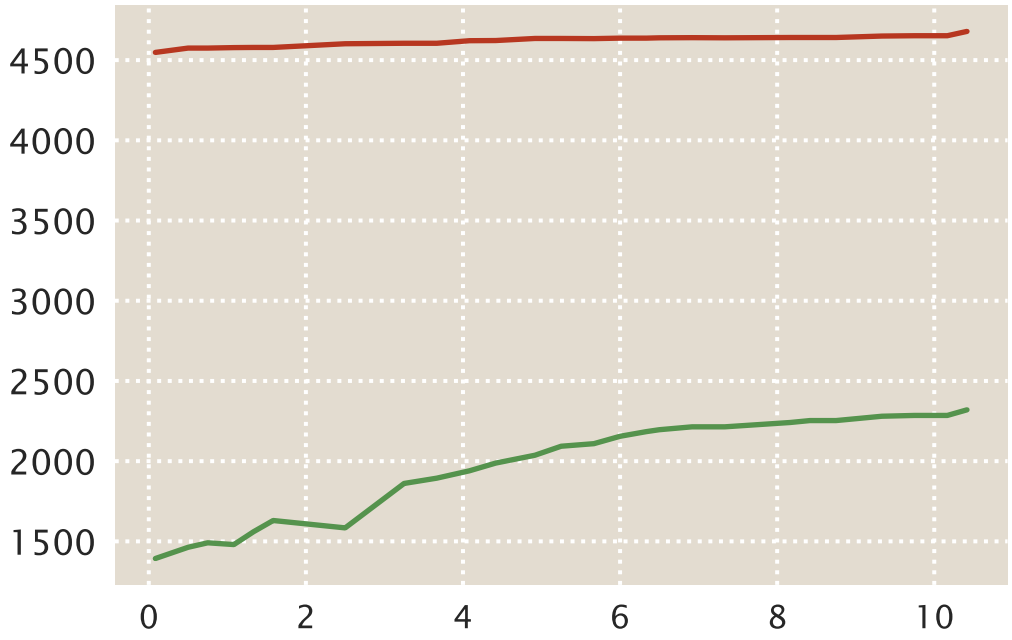

Figure 1.8: **Lack of regulatory annotation in *E. coli*.** (A) Operons in the *E. coli* genome are color-coded according to whether they have regulatory annotation (blue) or no regulatory annotation (red) in RegulonDB. (B) The relative number of recorded genes (red) and transcription factor binding sites (green) are plotted for each RegulonDB release over the past 10 years.

1.6 Experimental methods for discovering and analyzing regulatory motifs

A central goal of the present work is to dissect regulatory regions "from start to finish," which includes discovering and identifying transcripiton factor binding sites for unannotated promoters, determining the identities and regulatory roles of these transcription factors, and measuring transcription factor binding energies. Ultimately it is desirable to perform such an analysis in a high-throughput manner for unannotated promoters genome-wide, but this will require the development of new methodology. However, a number of methods currently exist that can be leveraged to dissect regulatory architectures on a low- to mid-throughput scale. Here we discuss three classes of methods that have been used previously to analyze regulatory architectures: occupancy assays (Figure 1.9A), *in vitro* affinity assays (Figure 1.9B), and massively parallel reporter assays (MPRAs) (Figure 1.9C). We give special attention to MPRAs as they are used extensively in this work.

Occupancy assays create maps of the locations of nucleotide-binding elements throughout the genome. For example, Chromatin ImmunoPrecipitation (ChIP)-seq is a common technique for determining the locations of transcription factors and histones [47, 48]. In a ChIP-seq experiment, the genome is fragmented and antibodies are introduced which target a transcription factor or histone of interest. These antibodies are immunoprecipitated to produce a sample containing the targeted protein along with any bound DNA fragments. These DNA fragments are sequenced and aligned to the genome to create a map of the binding sites for the targeted protein. Similar occupancy-based techniques can be used for applications such as determining the distribution of ribosomes [49] and identifying nucleosome binding regions, regions of open chromatin, and other regulatory elements in eukaryotes [50–54].

Occupancy assays can be used to determine the rough sequence specificity of a specific transcription factor, as they provide multiple examples of sequences that bind to the transcription factor of interest. However, these assays provide no information regarding transcription factor affinity–that is, the binding energy between the transcription factor and a given sequence. A number of *in vitro* methods have been devised to sensitively determine transcription factor sequence specificity and binding affinity [55–58]. *In vitro* methods allow one to assay the interactions between purified transcription factors and thousands of sequence variants. For example, protein-binding microarrays (PBMs) are a common, straightforward assay for assessing sequence specificity. In this assay, a microarray spotted with thousands of DNA sequence variants is incubated with fluorescently labeled transcription factor.

The transcription factor binds with some probability to each DNA spot, depending on the affinity of the transcription factor to the DNA sequence. Measuring the fluorescent intensity of each spot allows one to determine the affinity of the transcription factor for the DNA sequence within the spot [55]. Other *in vitro* methods such as MITOMI [56], HT-SELEX [57], and Spec-seq [58] similarly allow for high-sensitivity measurements of transcription factor binding affinities and sequence specificities. A distinct advantage of *in vitro* techinques is that they can be used to analyze low-affinity binding events [56, 59, 60]. However, a major drawback of *in vitro* techniques is that they cannot fully capture the subtleties of *in vivo* protein binding, which includes competition from other proteins, the influence of small molecules, and DNA shape effects. Additionally, both *in vitro* methods and occupancy methods focus on specific proteins that must be purified or immunoprecipitated, and thus are not especially useful for analyzing the regulatory architectures of specific promoters which may lack full regulatory annotation.

Massively parallel reporter assays (MPRAs, reviewed in Refs. [61, 62] and schematized in Figure 1.9C) are a diverse, versatile class of assays that can be used to analyze multiple aspects of transcriptional regulation either locally or genome-wide. In general, MPRAs are performed by positioning a library of promoter variants upstream of a reporter gene. Variations to the promoter can include single-nucleotide mutations to the promoter region [13, 63–65], transcription factor arrangement [66–68], spacing between binding sites [66, 68], or any other modification that can be made at the nucleotide level. These variations alter the promoter's regulatory properties, resulting in a change in the reporter gene's expression level. If the reporter gene is fluorescent, the cells may then be sorted into bins according to their fluorescence using fluorescence-activated cell sorting (FACS) [13, 66]. The contents of each bin are sequenced, and the sequence of each promoter variant is thereby associated with the reporter gene's expression level. Another common strategy is to associate the promoter variant with a barcode that is transcribed along with the reporter gene [63–65, 67, 68]. One can then sequence the reporter gene's mRNA transcripts and count the number of times that each barcode appears, thus associating gene expression with promoter variant in a fine-grained manner.

A number of studies use MPRAs to assay systematic perturbations to individual promoters [13, 63–68]. In addition, many variations on MPRAs have been developed to assay numerous other aspects of transcriptional regulation, such as analyzing regulatory "parts" for synthetic biology applications [69] or testing models for

predicting regulatory motifs in human cells [70, 71]. With minor modifications, the technique is also well-suited to genome-wide analysis and discovery of transcription factor binding sites [72–77]. Additionally, MPRAs can be combined with other techniques to obtain a more detailed understanding of regulatory systems. For example, MPRAs have been combined with occupancy assays to identify candidate enhancers and correlate transcription factor occupancy with regulation [78–80].

In this work we make use of the MPRA Sort-Seq [13] to discover transcription factor binding sites, infer the sites' regulatory roles and interactions, and create predictive models of transcription factor-DNA binding. We innovate by showing that we can thoroughly dissect promoters with diverse regulatory mechanisms given little to no initial information regarding a promoter's regulation. Additionally, we show that when combined with the proper analysis techniques, Sort-Seq can be used to determine transcription factor binding affinity with accuracy comparable to *in vitro* assays.

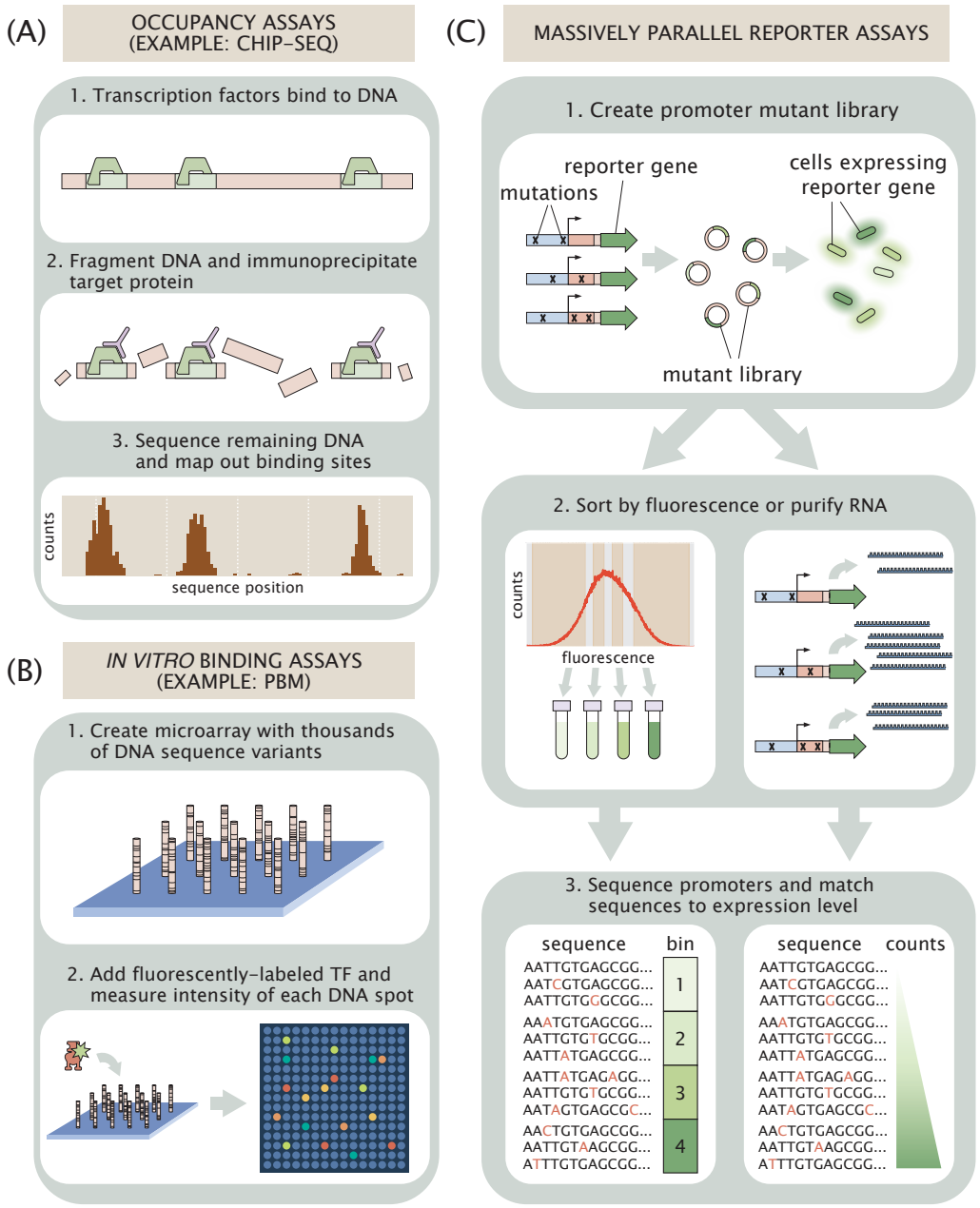

Figure 1.9: **Diverse methods assay multiple aspects of transcription factor binding.** (A) Occupancy-based methods can determine the locations of regulatory elements throughout the genome. For example, ChIP-seq works by digesting DNA and then immunoprecipitating a transcription factor of interest. Sequencing the DNA fragments attached to the transcription factor and aligning these fragments to the genome provides a map of transcription factor binding sites. (B) *In vitro* assays provide highly accurate readings of transcription factor sequence specificity and affinity. For example, protein binding microarrays (PBMs) contain thousands of DNA sequences to which fluorescently-labeled transcription factors can bind. The fluorescent intensity of each DNA spot on the microarray then serves as a readout of the transcription factor's affinity for the associated DNA sequence. (C) MPRAs are performed by positioning a library of promoter variants upstream of a reporter gene. Measuring the expression of the reporter gene and correlating expression with promoter sequence makes it possible to ascertain the roles of promoter elements.

1.7 Computational methods for analyzing data

This work focuses on using quantitative thinking to interpret and predict biological phenomena. A key component of a theory-experiment dialogue is the ability to analyze data using appropriate computational methods. The question of which computational methods are "appropriate" is a deep subject that goes beyond the scope of this work (consider, for instance, the current debate about p-values and the reproducibility of results [81]). In general we seek to analyze our data in ways that allow us to honestly assess the plausibility of our hypotheses. Additionally, we seek analysis methods that allow us to draw deeper inferences from our data, as in Chapter 4 where we apply inference methods that allow us to use massively parallel sequencing data to analyze the energetics of transcription factor binding in addition to the simpler task of identifying sequence preferences. Throughout this work, we make extensive use of methods for inferring parameter values and assessing the significance of experimental results. Here we discuss some key concepts that inform these methods: Bayesian inference, Markov Chain Monte Carlo, and mutual information.

The basics of Bayesian inference

Parameter fitting plays an important role in this work. Fitting can be accomplished using a number of methods, such as least squares fitting or linear regression. In general, the fitting methods used throughout this work rely on Bayesian inference (see Sivia and Skilling, Ref. [82] for a thorough treatment on the subject). Bayesian inference uses the premise that any hypothesis has a probability associated with it that represents one's certainty about the hypothesis. In the context of parameter fitting, the hypothesis is a proposed value for the parameter of interest. The probability that the proposed value is correct will change as one learns more about the system, for example by gathering data.

To provide an example of how this premise is used to infer parameter values, we can consider the example of a transcription factor binding site with an unknown binding energy, $\Delta\varepsilon_{tf}$. We wish to determine the value of $\Delta\varepsilon_{tf}$. Given no additional information about the binding site, but knowing a little bit about how transcription factor binding sites work, we may be able to make a rough guess about the value of $\Delta\varepsilon_{tf}$ in the form of a probability distribution, $P(\Delta\varepsilon_{tf})$. We know that $\Delta\varepsilon_{tf}$ must be less than 0 k_BT, otherwise it would not be energetically favorable for a transcription factor to bind to the site. We can also assume that $\Delta\varepsilon_{tf}$ must be greater than -20 k_BT, which provides a buffer around the approximate theoretical limit of transcription

factor binding energy of $\sim 15\ k_BT$ [83]. If we make no further assumptions, we can represent our guess about the value of $\Delta\varepsilon_R$ as a uniform distribution given by

$$P(\Delta\varepsilon_{tf}) = \begin{cases} \frac{1}{20} & -20\ k_BT \leq \Delta\varepsilon_{tf} \leq 0\ k_BT \\ 0 & \text{otherwise} \end{cases}. \tag{1.22}$$

In the language of Bayesian inference, we can refer to $P(\Delta\varepsilon_{tf})$ as the "prior," as it reflects our knowledge of the parameter value prior to learning any additional information about the system.

The prior distribution doesn't do much to help us determine the value of $\Delta\varepsilon_{tf}$. This is where we need to use the Bayesian concept of updating our hypothesis with new information. For example, let's say we can go in and directly measure the binding energy of the binding site. As we'll see in Chapter 4, determining transcription factor binding energies isn't as simple as just taking a direct measurement, but for the purpose of illustration we will imagine that we have some kind of nanoscale "energy meter" that can take these sorts of measurements. This hypothetical set-up is illustrated in Figure 1.10A. Even if we could measure the binding energy in this way, each measurement would have some amount of error associated with it. Therefore we should take some number of measurements to build up a data set, D, which will allow us to infer a more accurate value for the binding energy. This allows us to update our prior probability distribution $P(\Delta\varepsilon_{tf})$ (shown in Figure 1.10B) given our data D, giving us the posterior probability distribution $P(\Delta\varepsilon_{tf}|D)$. This term is referred to as the "posterior," as it reflects our knowledge of the parameter value after we learn more information about the system.

How do we come up with an expression for $P(\Delta\varepsilon_{tf}|D)$? Bayesian statistics provides us a solution in the form of Bayes' theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1.23}$$

Rewriting this for the example of our transcription factor binding site gives us

$$P(\Delta\varepsilon_{tf}|D) = \frac{P(D|\Delta\varepsilon_{tf})P(\Delta\varepsilon_{tf})}{P(D)}. \tag{1.24}$$

We have already discussed $P(\Delta\varepsilon_{tf}|D)$ and $P(\Delta\varepsilon_{tf})$, but the terms $P(D|\Delta\varepsilon_{tf})$ and $P(D)$ are new. $P(D|\Delta\varepsilon_{tf})$ is the probability that our data set could occur given

some proposed value for $\Delta\varepsilon_{tf}$. This term is referred to the "likelihood," and it will be discussed further below. $P(D)$ is the probability of the data set occurring at all, which can be difficult to determine, but fortunately we can treat it as a proportionality constant, giving us

$$P(\Delta\varepsilon_{tf}|D) \propto P(D|\Delta\varepsilon_{tf})P(\Delta\varepsilon_{tf}). \tag{1.25}$$

This works because we are only interested in the relative probabilities of different values for $\Delta\varepsilon_{tf}$. Now, in order to update our hypothesis for the value of $\Delta\varepsilon_{tf}$ based on our data set D, we need to find an expression for the likelihood, $P(D|\Delta\varepsilon_{tf})$. This is simple enough in our hypothetical scenario in which we can directly measure the transcription factor binding energy, as we can assume that our energy readings are normally distributed about the true energy value, $\Delta\varepsilon_{tf}$. Each data point x then has the associated probability distribution

$$P(x|\Delta\varepsilon_{tf}, \sigma) = \frac{1}{\sqrt{(2\pi\sigma)}} \exp\left[-\frac{(x - \Delta\varepsilon_{tf})^2}{2\sigma^2}\right], \tag{1.26}$$

where the error in the measurement x is represented by the quantity $(x - \Delta\varepsilon_{tf})$. We also note that we have introduced a new parameter, σ, which represents the standard deviation of the normal distribution. As will be discussed later, in many situations we will not know the value of σ, and there are ways to get around this. However, for the purposes of this illustration, we will assume that we know our hypothetical "energy meter" has an accuracy of $\pm 1.0\ k_BT$, so that $\sigma = 1.0$.

Now we have a probability distribution for each individual data point, but we want an expression for $P(D|\Delta\varepsilon_{tf})$, where D is a data set containing multiple energy readings. If we assume that each energy reading is independent, then $P(D|\Delta\varepsilon_{tf})$ is just the product of the probability for each individual reading in D, giving us

$$P(D|\Delta\varepsilon_{tf}, \sigma) = \prod_i^N \frac{1}{\sqrt{(2\pi\sigma)}} \exp\left[-\frac{(x_i - \Delta\varepsilon_{tf})^2}{2\sigma^2}\right] = (2\pi\sigma)^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_i^N (x_i - \Delta\varepsilon_{tf})^2\right], \tag{1.27}$$

where x_i is the ith measurement of N total measurements.

Now that we have expressions for the prior $P(\Delta\varepsilon_{tf})$ and the likelihood $P(D|\Delta\varepsilon_{tf})$, we can write an expression for $P(\Delta\varepsilon_{tf}|D)$,

$$P(\Delta\varepsilon_{tf}|D) \propto \begin{cases} \frac{1}{20}(2\pi\sigma)^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_i^N (x_i - \Delta\varepsilon_{tf})^2\right] & -20\ k_BT \leq \Delta\varepsilon_{tf} \leq 0\ k_BT \\ 0 & \text{otherwise} \end{cases}. \tag{1.28}$$

We can simplify this quite a bit. Since we are working with proportionalities, and since we can expect the normal distribution to go to zero at high or low values of $\Delta\varepsilon_{tf}$, we can work with a simplified expression for the posterior,

$$P(\Delta\varepsilon_{tf}|D) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_i^N (x_i - \Delta\varepsilon_{tf})^2\right]. \tag{1.29}$$

The posterior probability distribution $P(\Delta\varepsilon_{tf}|D)$ is plotted in Figure 1.10C for several example data sets with different numbers of data points, N. Each curve is normalized so that it sums to 1. The true value of $\Delta\varepsilon_{tf}$ that was used to generate these curves is $\Delta\varepsilon_{tf} = -14k_BT$, represented by the gray dotted line in Figure 1.10C. We see that larger data sets give us better estimates of this binding energy, and they also result in much narrower distributions, which reflects the amount of certainty associated with the estimate.

Of course, we cannot actually directly measure transcription factor binding energy using a nanoscale energy meter. In order to actually determine transcription factor binding energies, one must devise experiments that make it possible to infer the binding energy based on some other measurement. This type of inference is a central element of Chapter 4, wherein we create simple repression constructs with repressor binding sites of unknown binding energies, and take fold-change measurements that allow us to infer the energies. We compare our fold-change measurements to a theoretical value for the fold-change given by Equation 1.15, which we denote as $fc(\Delta\varepsilon_R, R)$. This gives us a value for the error in our measurement given a proposed value of $\Delta\varepsilon_R$ and a known value of R. This results in a scenario quite similar to the hypothetical one outlined in Figure 1.10. We are again dealing with Gaussian-distributed error, which allows us to assume a Gaussian distribution for the likelihood function and a uniform distribution for the prior. Unlike our hypothetical example, however, we do not know the value of σ. As has been covered in detail

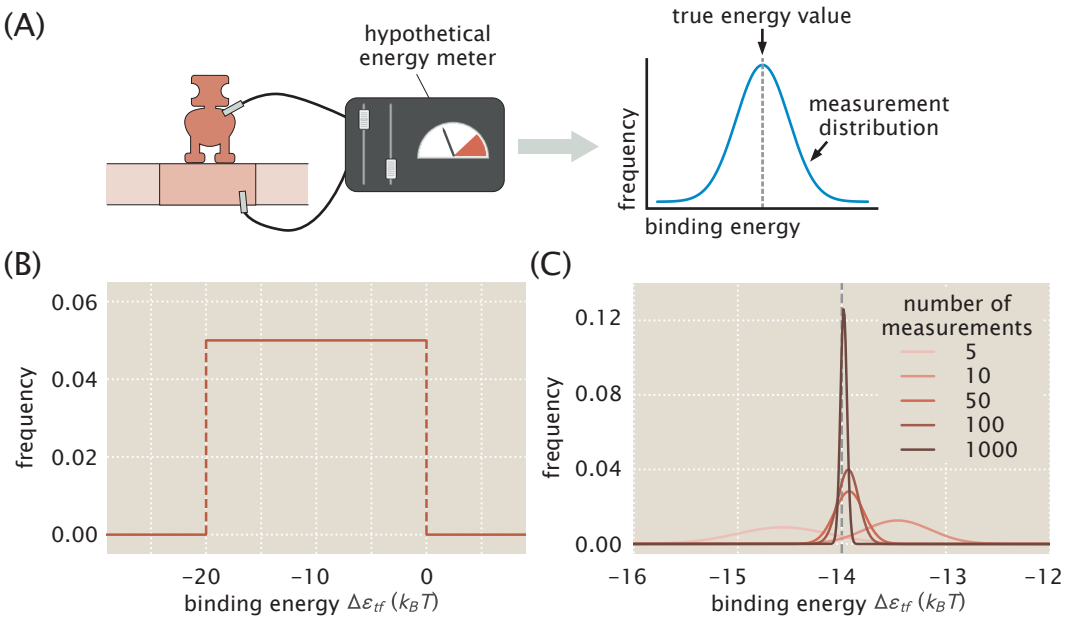

Figure 1.10: **Using Bayesian inference to determine the value of a parameter.** (A) For the purposes of illustration, we imagine a transcription factor and binding site with unknown binding energy $\Delta\varepsilon_{tf}$. We are able to measure the binding energy using a hypothetical nanoscale energy meter. However, the energy meter is not perfectly accurate, and takes measurements that are normally distributed about the true value of $\Delta\varepsilon_{tf}$. (B) Before taking any measurements, we can make a guess as to what the transcription factor binding energy may be. This is our prior, notated here as $P(\Delta\varepsilon_{tf})$. This represents the information we know about a parameter before taking data. For this system, our prior is uniform within the range of allowed transcription factor binding energies, and 0 everywhere else. (C) After taking measurements, we can update our knowledge about the value of $\Delta\varepsilon_{tf}$ given the data to obtain $P(\Delta\varepsilon_{tf}|D)$. This is our posterior distribution for the value of $\Delta\varepsilon_{tf}$. The shape and position of the posterior varies depending on the the data set that is used to construct it. This reflects the fact that both accuracy and certainty are improved by taking more data. The true value of $\Delta\varepsilon_{tf}$ is illustrated by the vertical dotted line in the plot.

elsewhere [82], under these conditions the posterior distribution can be represented by a student-t distribution, which for our simple repression system is given by

$$P(\Delta\varepsilon_R|D) \propto \left[\sum_i^N \left(fc_{i,exp} - fc(\Delta\varepsilon_R, R_i)\right)^2\right]^{\frac{N}{2}}, \tag{1.30}$$

where $fc_{i,exp}$ is the value of the ith data point and R_i is the value of R used to obtain that data point. In Chapter 4, we compute this distribution using theoretical fold-change values that are calculated using an array of values for $\Delta\varepsilon_R$. In this way,

we fit the fold-change equation to the fold-change data in order to find the most probable value of $\Delta\varepsilon_R$ for the repressor binding site.

The basics of Markov Chain Monte Carlo

In the previous section, we provided examples of parameter estimation by Bayesian inference. The expressions we worked with in that section were relatively simple and could be manipulated analytically. They also had a very straightforward interpretation: the parameter value that produced the largest posterior probability is the most likely parameter value.

However, we will often want to infer parameter values under conditions that are not so tidy. The posterior distribution may be difficult or impossible to work with analytically, or we may wish to infer many parameters at once, some of which might be correlated with one another. Chapters 2 through 4 contain examples of models that present some of these difficulties. In Chapter 2, in which we implement predictive models of allosteric simple repression, we wish to infer two or more parameters at once, which is complicated by some troublesome correlations between parameters. Chapters 3 and 4 involve modeling transcription factor binding sites using energy matrices, which require us to use Sort-Seq data to infer $4 \times L$ parameters at once, where L is the length of the transcription factor binding site. These examples require a more powerful method than simply plotting posterior distributions and finding the parameter value that maximizes the posterior.

A method that is commonly used for these types of scenarios is Markov Chain Monte Carlo, or MCMC. MCMC is a deep topic that lends itself to highly technical discussions. Here we aim to explain the basics of MCMC in straightforward terms, following the example set forth in Ref. [84].

MCMC gets its name from two processes, Monte Carlo and Markov Chain. Monte Carlo is a method for estimating features of a distribution by randomly drawing samples from the distribution. For example, one could estimate the mean or standard deviation of a distribution by drawing random samples and computing the mean and standard deviation for those samples. Figure 1.11A schematizes this process. Markov Chain is a type of algorithm for drawing random samples. In a Markov Chain, schematized in Figure 1.11B, some calculation is performed on each sample to generate the next sample. While each sample relies on the previous sample, however, there is no memory of any samples before the previous one. When we put these concepts together in MCMC, we get a randomly-generated chain of samples that can be used to approximate the properties of a posterior distribution. In fact, the histogram of the values that make up the chain should approximately reproduce the posterior distribution.

The process of generating random samples in a way that will approximate a posterior distribution is the topic of much study, and there are a number of algorithms that are used to perform this process depending on the scenario [85]. One of the simplest and most common algorithms is the Metropolis-Hastings algorithm. The algorithm begins by choosing a starting trial parameter value, if possible one that is believed to be close to the true parameter value. Then, a new trial is generated by adding a random "jump" parameter to the current value. The jump parameter is drawn from a symmetric distribution that is centered at zero. The new trial will now either be accepted and added to the chain or rejected and discarded. If the trial is discarded, then the next value in the chain is a copy of the previous value. If the posterior probability calculated with the new trial is higher than the posterior probability calculated with the previous value, then the new trial is accepted. If not, then the new trial is accepted with a probability equal to $\frac{P(\mu_{\text{new}}|D)}{P(\mu_{\text{old}}|D)}$, where μ is the proposed parameter value. Occasionally accepting trials with lower posterior probabilities than the previous value allows the chain to avoid getting stuck in local maxima. This algorithm is repeated until a specified number of iterations is reached. After completing MCMC using the Metropolis-Hastings algorithm, the histogram of the samples in the chain should approximate the posterior distribution from which you were drawing the samples.

To provide a simple example of MCMC in action, we now will use it to approximate the binding energy of a repressor binding site. We use data from Chapter 4 in which fold-change measurements were taken for simple repression constructs using this binding site. These measurements were taken using multiple background strains with varying known repressor copy numbers R. While MCMC is not necessary to fit for the value of $\Delta\varepsilon_R$ for this binding site, and indeed was not used to perform this fit in Chapter 4, we use MCMC here to provide a conceptually straightforward example. As mentioned earlier, the posterior distribution for the transcription factor binding energy can be represented by the student-t distribution given in Equation 1.30. To estimate the value of $\Delta\varepsilon_R$ given a set of fold-change data, we use the Metropolis-Hastings algorithm, plugging the proposed value of $\Delta\varepsilon_R$ and our set of experimental fold-change measurements D into Equation 1.30 to evaluate the posterior probability for each trial, and then add trials to the chain according to the rules of the algorithm.

As an example of what MCMC chains look like, a short 100-iteration chain estimating the binding energy of a repressor binding site is shown in Figure 1.11C. We note

that MCMC chains are generally much longer than this, and the short length was chosen so that each step in the chain could be visually distinguished in the figure. The chain shown in Figure 1.11C started with a trial of $\Delta\varepsilon_R = -12\ k_BT$, which is close to the fitted value of $\Delta\varepsilon_R = -12.24\ k_BT$ for this binding site. Because the starting trial is close to the fitted value, the chain never strays far from the fitted value. The histogram of the values in the chain, plotted in Figure 1.11D, approximates the actual posterior distribution for $\Delta\varepsilon_R$, shown as a dotted line.

In actual practice, MCMC chains will be much longer than 100 iterations. This is partially due to the fact that estimates improve as the number of samples increases. Another reason that it is important to do many iterations is a phenomenon known as "burn-in." Burn-in relates to the dilemma of choosing a good initial trial value. If possible, it is a good idea to start with the maximum likelihood estimate for the parameter value. Depending on how much is known about the system, however, it may not be possible to reliably choose an initial parameter trial that is close to the actual parameter value. If the initial parameter value is sufficiently distant from the actual value, the chain may search parameter space for some time before coming close enough to the actual value for the chain to converge around the actual value. This search is known as the burn-in period. Figure 1.11E illustrates the burn-in period for the case of our repressor binding site with $\Delta\varepsilon_R = -12.24\ k_BT$. When our initial parameter value is $-12\ k_BT$ there is no burn-in to speak of, and the chain immediately converges to a distribution around $-12.24\ k_BT$. However, if we choose an initial parameter value of $-15\ k_BT$ there is a noticeable burn-in period, and if we stray farther from the true binding energy with an initial parameter value of $-18\ k_BT$, the burn-in period is even longer. The elements of the chain associated with the burn-in period should not be included when using the chain to estimate parameter values. For situations where it is difficult to begin with a good estimate for the parameter value, it is common to run multiple MCMC chains with a variety of initial trial values to increase the chances of finding a chain that quickly converges around the true parameter value.

Here we have provided a basic primer regarding the use of MCMC for parameter estimation. We go into greater detail in Chapters 3 and 4 regarding how MCMC was implemented in our work.

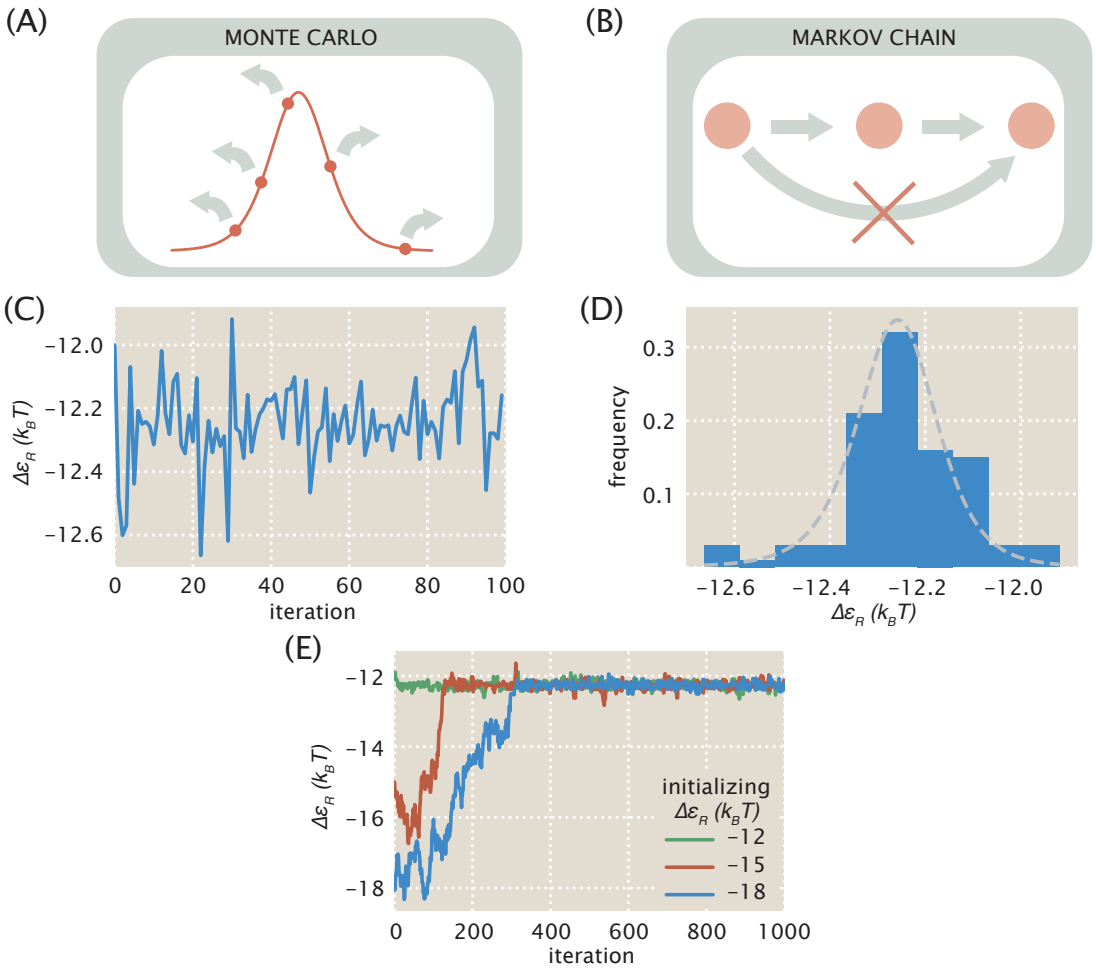

Figure 1.11: **Parameter inference using Markov Chain Monte Carlo.** (A) "Monte Carlo" refers to the process of taking random samples from a distribution and using these samples to estimate properties of the distribution. (B) "Markov Chain" refers to a process of generating random samples in which an operation is performed on the current sample to generate the next sample. The identity of each sample depends only on the sample immediately preceding it, and no other previous samples. (C) A short MCMC chain was generated using a student-t posterior distribution and fold-change data from a simple repression construct from Chapter 4. The chain was formed by generating trial values for the repressor binding energy, $\Delta\varepsilon_R$, using a Metropolis-Hastings algorithm. (D) A histogram generated from the chain plotted in (C) shows that the chain approximates a student-t posterior distribution (dotted line) centered at the best estimate for the repressor binding energy, $\Delta\varepsilon_R = -12.24\ k_BT$. (E) Initiating an MCMC chain with different trial values can cause the chain to have a "burn-in" period in which it searches parameter space before converging around the best estimate for the parameter.

Mutual information

In a genome's coding regions, it is clear that DNA contains a wealth of information in the form of instructions for building proteins. Although there is no distinct code connecting DNA sequence to regulatory activity, the noncoding DNA in the promoter region is information-rich as well. In this work we borrow tools from information theory to help interpret the relationship between DNA sequence and transcriptional regulation.

The two key information theory concepts used in this work are Shannon entropy and mutual information. The Shannon entropy is given by the equation

$$S(p_i) = -\sum_{i}^{N} p_i \log_2 p_i, \tag{1.31}$$

where p_i is the probability of the ith microstate or outcome available to the system. The Shannon entropy, or "missing information," reflects our level of uncertainty about the state of a system. To show how Shannon entropy quantifies uncertainty, we consider the example of a fair coin versus a biased coin. Both a fair coin and a biased coin have two states available to them: heads or tails. For a fair coin, $p_{\text{heads}} = p_{\text{tails}} = 0.5$. For a biased coin, however, $p_{\text{heads}} \neq p_{\text{tails}}$. For this example let's say that for the biased coin $p_{\text{heads}} = 0.75$ and $p_{\text{tails}} = 0.25$. Now if we flip a coin, how certain are we that it will land heads? For the biased coin we are more certain that it will land heads than we are for the fair coin. To quantify this certainty using Shannon entropy we can calculate

$$S_{\text{fair}}(p_i) = -\left(0.5\log_2 0.5 + 0.5\log_2 0.5\right) = 1.0 \text{ bits} \tag{1.32}$$

and

$$S_{\text{biased}}(p_i) = -\left(0.75\log_2 0.75 + 0.25\log_2 0.25\right) \approx 0.81 \text{ bits}. \tag{1.33}$$

Here we use the unit "bits" to quantify information, as will be discussed further below. We can see that $S_{\text{fair}}(p_i) > S_{\text{biased}}(p_i)$. This indicates that we have greater *uncertainty* regarding the outcome of a coin flip using a fair coin than a biased coin.

How does measuring our uncertainty allow us to relate DNA sequence to gene regulation? This becomes clearer when we introduce the idea of mutual information.

Mutual information quantifies the relatedness between two quantities. Specifically, it quantifies the extent to which knowing the value of parameter B reduces your uncertainty regarding the value of parameter A. Mutual information can be written in terms of Shannon entropy as

$$I(A;B) = S(A) - S(A|B). \tag{1.34}$$

When we employ the definition of Shannon entropy given in Equation 1.33, this evaluates to

$$I(A;B) = \sum_{i,\,j}^{N} P(A_i, B_j) \log_2 \left(\frac{P(A_i, B_j)}{P(A_i)P(B_j)} \right). \tag{1.35}$$

We can use mutual information to quantify the extent to which a given regulatory sequence element (A) contributes to the level of gene expression (B). As an example of how this works, we consider a hypothetical dinucleotide placed upstream of a RNAP binding site, as schematized in Figure 1.12A. For a variety of dinucleotide sequence combinations we measure the gene expression associated with this promoter. We wish to determine the extent to which the identities of the bases in the dinucleotide influence the level of gene expression, which we can do using Equation 1.35. To keep this example simple, we consider only whether each base in the dinucleotide is a purine or a pyrimidine (denoted R or Y, respectively) and whether the gene expression is high or low (see Figure 1.12B). We outline all relevant probabilities in Figure 1.12C. We can plug the values for position 1 (indicated as a subscript in the equation below) into Equation 1.35 to get

$$\begin{aligned}
I(A_1;B) &= P(R_1, \text{high}) \log_2 \left(\frac{P(R_1, \text{high})}{P(R_1) \times P(\text{high})} \right) + P(R_1, \text{low}) \log_2 \left(\frac{P(R_1, \text{low})}{P(R_1) \times P(\text{low})} \right) \\
&\quad + P(Y_1, \text{high}) \log_2 \left(\frac{P(Y_1, \text{high})}{P(Y_1) \times P(\text{high})} \right) + P(Y_1, \text{low}) \log_2 \left(\frac{P(Y_1, \text{low})}{P(Y_1) \times P(\text{low})} \right) \\
&= 0.25 \log_2 \left(\frac{0.25}{0.5 \times 0.5} \right) + 0.25 \log_2 \left(\frac{0.25}{0.5 \times 0.5} \right) \\
&\quad + 0.25 \log_2 \left(\frac{0.25}{0.5 \times 0.5} \right) + 0.25 \log_2 \left(\frac{0.25}{0.5 \times 0.5} \right) \\
&= \textbf{0 bits}.
\end{aligned} \tag{1.36}$$

Next we plug the values for position 2 (indicated as a subscript in the equation below) into Equation 1.35 to get

$$
\begin{aligned}
I(A_2;B) &= P(R_2,\text{high})\log_2\left(\frac{P(R_2,\text{high})}{P(R_2)\times P(\text{high})}\right) + P(R_2,\text{low})\log_2\left(\frac{P(R_2,\text{low})}{P(R_2)\times P(\text{low})}\right) \\
&\quad + P(Y_2,\text{high})\log_2\left(\frac{P(Y_2,\text{high})}{P(Y_2)\times P(\text{high})}\right) + P(Y_2,\text{low})\log_2\left(\frac{P(Y_2,\text{low})}{P(Y_2)\times P(\text{low})}\right) \\
&= 0\log_2\left(\frac{0}{0.5\times 0.5}\right) + 0.5\log_2\left(\frac{0.5}{0.5\times 0.5}\right) \\
&\quad + 0.5\log_2\left(\frac{0.5}{0.5\times 0.5}\right) + 0\log_2\left(\frac{0}{0.5\times 0.5}\right) \\
&= \textbf{1.0 bits}. \qquad (1.37)
\end{aligned}
$$

We find that $I(A_1;B) = 0$ bits and $I(A_2;B) = 1$ bit. These are the minimum and maximum possible mutual information values for this system, respectively. This means that the expression level for our hypothetical promoter is entirely determined by the identity of the base at position 2. The use of the information unit "bit" references the number of yes/no questions regarding A that are required in order to determine the value of B. In this case, one yes/no question is required (i.e., is A a purine?) so the maximum mutual information between parameters is 1 bit.

In actual practice in Chapters 3 and 4, we look at much longer segments of regulatory DNA, consider all four base identities for each sequence position, and measure four different levels of gene expression. The mutual information between sequence and expression in these experiments is also much less definitive than in the toy example given in Figure 1.12A-C. We rarely, if ever, see situations in which the identity of a single nucleotide is the sole determinant of a promoter's expression level. However, as discussed in detail in Chapter 3, mutual information serves as a useful tool for identifying the relative importance of each nucleotide in a promoter region, which makes it possible to identify the locations of binding sites. Additionally, as discussed in Chapters 3 and 4, it provides a vital metric for inferring energy matrices and sequence logos. In Figure 1.12D we show examples from Chapter 3. Here, the mutual information between base identity and gene expression is calculated for the *relBE* promoter. Regions of high mutual information indicate possible RNAP or transcription factor binding sites. Additionally, energy matrices for binding sites are inferred using an MCMC technique in which the mutual information between predicted and measured gene expression values is maximized (see supplemental sections of Chapters 3 and 4 for more details). These principles are addressed in great detail in refs [13, 86].

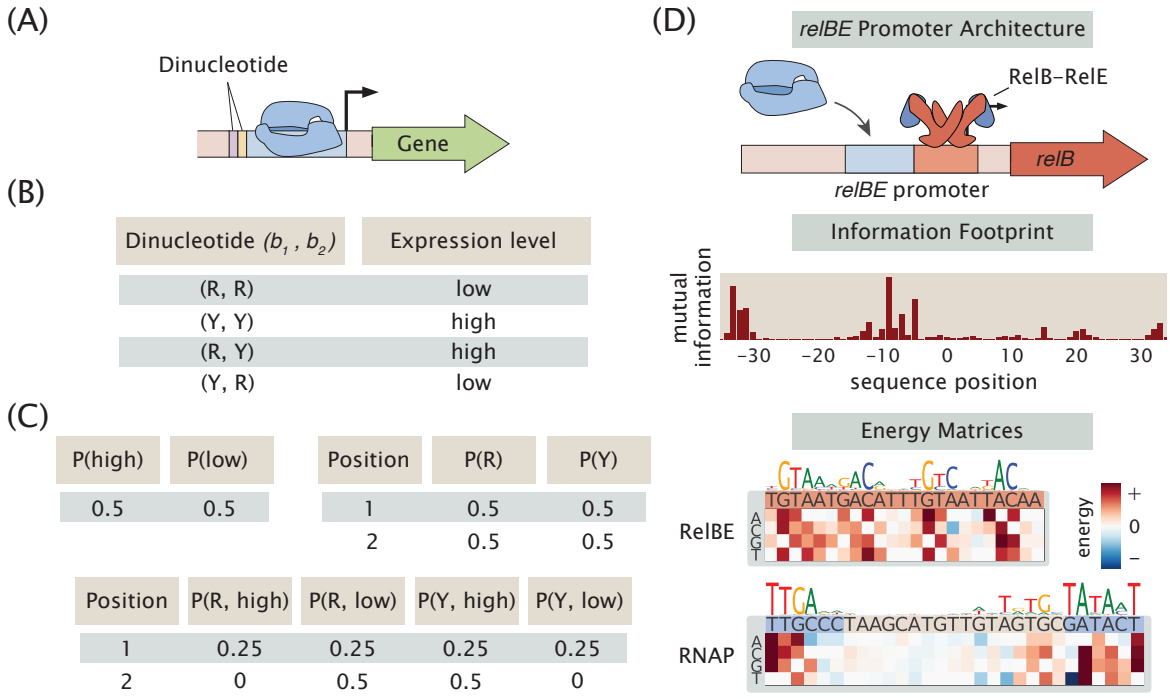

Figure 1.12: **Mutual information quantifies the relatedness of parameters.** (A) For the purposes of illustration, we consider a simplified promoter in which a regulatory dinucleotide is placed immediately upstream of the RNAP binding site. The identities of the bases in this dinucleotide are related to the expression level of the promoter. (B) We list some hypothetical data points in which we sequence the bases in the dinucleotide (which are labeled either as purines (R) or pyrimidines (Y)) and note whether the promoter's expression is high or low for each sequence. (C) We list the individual and joint probabilities associated with each base identity and expression level. These are used to calculate the mutual information between base identity and expression in the main text. (D) In Chapter 3 we use mutual information as a tool in characterizing regulatory architectures. Here we plot an "information footprint" in which we quantify the mutual information between base identity and gene expression for the *relBE* promoter. We also show the energy matrices for the RelBE and RNAP binding sites, which were inferred using mutual information.

BIBLIOGRAPHY

[1] Alan E. Guttmacher and Francis S. Collins. Welcome to the genomic era. *The New England Journal of Medicine*, 349(10):996–998, 2003.

[2] Lacramioara Bintu, Nicolas E. Buchler, Hernan G. Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips. Transcriptional regulation by the numbers: Models. *Current Opinion in Genetics and Development*, 15(2):116–124, 2005.

[3] H. G. Garcia and R. Phillips. Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences*, 108(29):12173–12178, 2011.

[4] Alvaro Sanchez, Hernan G. Garcia, Daniel Jones, Rob Phillips, and Jané Kondev. Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Computational Biology*, 7(3), 2011.

[5] Robert C. Brewster, Daniel L. Jones, and Rob Phillips. Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Computational Biology*, 8(12), 2012.

[6] Hernan G. Garcia, Alvaro Sanchez, James Q. Boedicker, Melisa Osborne, Jeff Gelles, Jane Kondev, and Rob Phillips. Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. *Cell Reports*, 2(1):150–161, 2012.

[7] James Q. Boedicker, Hernan G. Garcia, and Rob Phillips. Theoretical and experimental dissection of DNA loop-mediated repression. *Physical Review Letters*, 110(1):1–5, 2013.

[8] James Q. Boedicker, Hernan G. Garcia, Stephanie Johnson, and Rob Phillips. DNA sequence-dependent mechanics and protein-assisted bending in repressor-mediated loop formation. *Physical Biology*, 10(6):066005, 2013.

[9] Robert C. Brewster, Franz M. Weinert, Hernan G. Garcia, Dan Song, Mattias Rydenfelt, and Rob Phillips. The transcription factor titration effect dictates level of gene expression. *Cell*, 156(6):1312–1323, 2014.

[10] Mattias Rydenfelt, Hernan G. Garcia, Robert Sidney Cox, and Rob Phillips. The influence of promoter architectures and regulatory motifs on gene expression in *Escherichia coli*. *PLoS ONE*, 9(12):1–31, 2014.

[11] Franz M. Weinert, Robert C. Brewster, Mattias Rydenfelt, Rob Phillips, and Willem K. Kegel. Scaling of gene expression with transcription-factor fugacity. *Physical Review Letters*, 113(25):1–5, 2014.

[12] Jacques Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, 12:88–118, 1965.

[13] Justin B. Kinney, Anand Murugan, Curtis G. Callan Jr., and Edward C. Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*, 107(20):9158–9163, 2010.

[14] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[15] Patrick P. Dennis and Hans Bremer. Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus*, 3(1), 2008.

[16] Stefan Klumpp, Zhongge Zhang, and Terence Hwa. Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139(7):1366–1375, 2009.

[17] Martin Beck, Alexander Schmidt, Johan Malmstroem, Manfred Claassen, Alessandro Ori, Anna Szymborska, Franz Herzog, Oliver Rinner, Jan Ellenberg, and Ruedi Aebersold. The quantitative proteome of a human cell line. *Molecular Systems Biology*, 7(549):1–8, 2011.

[18] Samuel Marguerat, Alexander Schmidt, Sandra Codlin, Wei Chen, Ruedi Aebersold, and Jürg Bähler. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell*, 151(3):671–683, 2012.

[19] Alexander Schmidt, Karl Kochanowski, Silke Vedelaar, Erik Ahrné, Benjamin Volkmer, Luciano Callipo, Kèvin Knoops, Manuel Bauer, Ruedi Aebersold, and Matthias Heinemann. The quantitative and condition-dependent *Escherichia coli* proteome. *Nature Biotechnology*, 34(1):104–110, 2015.

[20] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3:318–356, 1961.

[21] Tony Romeo, Christopher A. Vakulskas, and Paul Babitzke. Post-transcriptional regulation on a global scale: Form and function of Csr/Rsm systems. *Environmental Microbiology*, 15(2):313–324, 2013.

[22] Susan Gottesman and Gisela Storz. Bacterial small RNA regulators: Versatile roles and rapidly evolving variations. *Cold Spring Harbor Perspectives in Biology*, 3(12):1–16, 2011.

[23] Luke E. Ulrich, Eugene V. Koonin, and Igor B. Zhulin. One-component systems dominate signal transduction in prokaryotes. *Trends in Microbiology*, 13(2):52–56, 2005.

[24] Gary K. Ackers, Alexander D. Johnson, and Madeline A. Shea. Quantitative model for gene regulation by lambda phage repressor. *Proceedings of the National Academy of Sciences*, 79:1129–1133, 1982.

[25] Rob Phillips, Jane Kondev, Julie Theriot, and Hernan Garcia. *Physical Biology of the Cell*. Garland Science, 2nd edition, 2012.

[26] Keiko Kimata, Hideyuki Takahashi, Toshifumi Inada, Pieter Postma, and Hiroji Aiba. cAMP receptor protein-cAMP plays a crucial role in glucose-lactose diauxie by activating the major glucose transporter gene in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 94:12914–12919, 1997.

[27] Robert Schleif. AraC protein, regulation of the l-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action. *FEMS Microbiology Reviews*, 34(5):779–796, 2010.

[28] Douglas F. Browning and Stephen J. W. Busby. Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, 14(10):638–650, 2016.

[29] Karsten Rippe, Peter H. von Hippel, and Jörg Langowski. Action at a distance: DNA-looping and initiation of transcription. *Trends in Biochemical Sciences*, 20:500–506, 1995.

[30] Axel Cournac and Jacqueline Plumbridge. DNA looping in prokaryotes: Experimental and theoretical approaches. *Journal of Bacteriology*, 195(6):1109–1119, 2013.

[31] Helmut Krämer, Monika Niemoller, Michèle Amouyal, Bernard Revet, Brigitte Von Wilcken-bergmann, and Benno Müller-hill. *lac* repressor forms loops with linear DNA carrying two suitably spaced *lac* operators. *EMBO Journal*, 6(5):1481–1491, 1987.

[32] Helmut Krämer, Michèle Amouyal, Alfred Nordheim, and Benno Müller-Hill. DNA supercoiling changes the spacing requirement of two *lac* operators for DNA loop formation with *lac* repressor. *EMBO Journal*, 7(2):547–556, 1988.

[33] Gregory R. Bellomy, Michael C. Mossing, and M. Thomas Record. Physical properties of DNA *in vivo* as probed by the length dependence of the *lac* operator looping process. *Biochemistry*, 27:3900–3906, 1988.

[34] Scott M. Law, Gregory R. Bellomy, Paula J. Schlax, and M. Thomas Record Jr. *In vivo* thermodynamic analysis of repression with and without looping in *lac* constructs. *Journal of Molecular Biology*, 230:161–173, 1992.

[35] Nicole A. Becker, Jason D. Kahn, and L. James Maher III. Bacterial repression loops require enhanced DNA flexibility. *Journal of Molecular Biology*, 349:716–730, 2005.

[36] Kevin B. Towles, John F. Beausang, Hernan G. Garcia, Rob Phillips, and Philip C. Nelson. First-principles calculation of DNA looping in tethered particle experiments. *Physical Biology*, 6, 2009.

[37] Stephanie Johnson, Yi Ju Chen, and Rob Phillips. Poly(dA:dT)-rich DNAs are highly flexible in the context of DNA looping. *PLoS ONE*, 8(10), 2013.

[38] Yi-Ju Chen, Stephanie Johnson, Peter Mulligan, Andrew J. Spakowitz, and Rob Phillips. Modulation of DNA loop lifetimes by the free energy of loop formation. *Proceedings of the National Academy of Sciences*, 111(49):17396–17401, 2014.

[39] Stephanie Johnson, Jan-Willem Van De Meent, Rob Phillips, Chris H. Wiggins, and Martin Lindén. Multiple LacI-mediated loops revealed by Bayesian statistics and tethered particle motion. *Nucleic Acids Research*, 42(16):10265–10277, 2014.

[40] Matthew Bush and Ray Dixon. The role of bacterial enhancer binding proteins as specialized activators of σ^{54}-dependent transcription. *Microbiology and Molecular Biology Reviews*, 76(3):497–529, 2012.

[41] Charles J. Dorman and Padraig Deighan. Regulation of gene expression by histone-like proteins in bacteria. *Current Opinion in Genetics and Development*, 13(2):179–184, 2003.

[42] Martijn S. Luijsterburg, Maarten C. Noom, G. J. L. Wuite, and Remus Th. Dame. The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: A molecular perspective. *Journal of Structural Biology*, 156(2):262–272, 2006.

[43] Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of DNA shape in protein-DNA recognition. *Nature*, 461(7268):1248–1253, 2009.

[44] Denise E. Waldron, Peter Owen, and Charles J. Dorman. Competitive interaction of the OxyR DNA-binding protein and the Dam methylase at the antigen 43 gene regulatory region in *Escherichia coli*. *Molecular Microbiology*, 44(2):509–520, 2002.

[45] Socorro Gama-castro, Heladia Salgado, Alberto Santos-zavaleta, Daniela Ledezma-tejeida, Luis Mu, Jair Santiago Garc, Kevin Alquicira-hern, Irma Mart, Lucia Pannier, Alejandra Medina-rivera, Hilda Solano-lira, Abraham Castro-mondrag, P. Ernesto, Shirley Alquicira-hern, L. Alejandra, Anastasia Hern, Del Moral-ch, Fabio Rinaldi, and Julio Collado-vides. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 44:133–143, 2016.

[46] Ingrid M. Keseler, Amanda Mackie, Alberto Santos-Zavaleta, Richard Billington, César Bonavides-Martínez, Ron Caspi, Carol Fulcher, Socorro Gama-Castro, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Luis Muñiz-Rascado, Quang Ong, Suzanne Paley, Martin Peralta-Gil, Pallavi Subhraveti, David A. Velázquez-Ramírez, Daniel Weaver, Julio Collado-Vides, Ian Paulsen, and Peter D. Karp. The EcoCyc database: Reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Research*, 45:D543–D550, 2017.

[47] David S. Johnson, Ali Mortazavi, and Richard M. Myers. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, 316:1497–1503, 2007.

[48] Peter J. Park. ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

[49] Nicholas T. Ingolia, Sina Ghaemmaghami, John R. S. Newman, and Jonathan S. Weissman. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science*, 324:218–223, 2009.

[50] G. E. Crawford, I. E. Holt, J. C. Mullikin, D. Tai, National Institutes of Health, R. Blakesley, G. Bouffard, A. Young, C. Masiello, E. D. Green, T. G. Wolfsberg, and F. S. Collins. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proceedings of the National Academy of Sciences*, 101(4):992–997, 2004.

[51] Paul G. Giresi, Jonghwan Kim, Ryan M. Mcdaniell, Vishwanath R. Iyer, and Jason D. Lieb. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, 17:877–885, 2007.

[52] Alan P. Boyle, Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132:311–322, 2008.

[53] Kristin Brogaard, Liqun Xi, Ji-Ping Wang, and Jonathan Widom. A map of nucleosome positions in yeast at base-pair resolution. *Nature*, 486:496–501, 2012.

[54] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nature Methods*, 10(12):1213–1218, 2013.

[55] Michael F. Berger and Martha L. Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature Protocols*, 4(3):393–411, 2009.

[56] Sebastian J. Maerkl and Stephen R. Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315:233–238, 2007.

[57] Yue Zhao, David Granas, and Gary D. Stormo. Inferring binding energies from selected binding sites. *PLoS Computational Biology*, 5(12), 2009.

[58] Zheng Zuo and Gary D. Stormo. High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. *Genetics*, 198(3):1329–1343, 2014.

[59] Daniel D. Le, Tyler C. Shimko, Arjun K. Aditham, Allison M. Keys, Yaron Orenstein, and Polly Fordyce. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proceedings of the National Academy of Sciences*, In press, 2018.

[60] Chaitanya Rastogi, H. Tomas Rube, Judith F. Kribelbauer, Justin Crocker, Ryan E. Loker, Gabriella D. Martini, Oleg Laptenko, William A. Freed-Pastor, Carol Prives, David L. Stern, Richard S. Mann, and Harmen J. Bussemaker. Accurate and sensitive quantification of protein-DNA binding affinity. *Proceedings of the National Academy of Sciences*, In press, 2018.

[61] Fumitaka Inoue and Nadav Ahituv. Decoding enhancers using massively parallel reporter assays. *Genomics*, 106(3):159–164, 2015.

[62] Michael A. White. Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. *Genomics*, 106:165–170, 2015.

[63] J. C. Kwasnieski, I. Mogno, C. A. Myers, J. C. Corbo, and B. A. Cohen. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the National Academy of Sciences*, 109(47):19498–19503, 2012.

[64] Alexandre Melnikov, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, Andreas Gnirke, Curtis G. Callan, Justin B. Kinney, Manolis Kellis, Eric S. Lander, and Tarjei S. Mikkelsen. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3):271–277, 2012.

[65] Rupali P. Patwardhan, Joseph B. Hiatt, Daniela M. Witten, Mee J. Kim, Robin P. Smith, Dalit May, Choli Lee, Jennifer M. Andrie, Su In Lee, Gregory M. Cooper, Nadav Ahituv, Len A. Pennacchio, and Jay Shendure. Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nature Biotechnology*, 30(3):265–270, 2012.

[66] Eilon Sharon, Yael Kalma, Ayala Sharp, Tali Raveh-Sadka, Michal Levo, Danny Zeevi, Leeat Keren, Zohar Yakhini, Adina Weinberger, and Eran Segal.

Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, 30(6):521–530, 2012.

[67] Ilaria Mogno, Jamie C. Kwasnieski, and Barak A. Cohen. Massively parallel synthetic promoter assays reveal the *in vivo* effects of binding site variants. *Genome Research*, 23:1908–1915, 2013.

[68] Robin P. Smith, Leila Taher, Rupali P. Patwardhan, Mee J. Kim, Fumitaka Inoue, Jay Shendure, Ivan Ovcharenko, and Nadav Ahituv. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics*, 45(9):1021–1028, 2013.

[69] Sriram Kosuri, Daniel B. Goodman, Guillaume Cambray, Vivek K. Mutalik, Yuan Gao, Adam P. Arkin, Drew Endy, and George M. Church. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 110(34):14024–14029, 2013.

[70] Pouya Kheradpour, Jason Ernst, Alexandre Melnikov, Peter Rogov, Li Wang, Jessica Alston, Tarjei S. Mikkelsen, and Manolis Kellis. Systematic dissection of motif instances using a massively parallel reporter assay. *Genome Research*, 23:800–811, 2013.

[71] Jamie C. Kwasnieski, Christopher Fiore, Hemangi G. Chaudhari, and Barak A. Cohen. High-throughput functional testing of ENCODE segmentation predictions. *Genome Research*, 24:1595–1602, 2014.

[72] Waseem Akhtar, Johann De Jong, Alexey V. Pindyurin, Ludo Pagie, Wouter Meuleman, Jeroen De Ridder, Anton Berns, Lodewyk F. A. Wessels, Maarten Van Lohuizen, and Bas Van Steensel. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell*, 154(4):914–927, 2013.

[73] Cosmas D. Arnold, Daniel Gerlach, Christoph Stelzer, L. M. Boryn, Martina Rath, Alexander Stark, Ł M Boryn, Martina Rath, Alexander Stark, L. M. Boryn, Martina Rath, and Alexander Stark. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123):1074–1077, 2013.

[74] Stephen S. Gisselbrecht, Luis A. Barrera, Martin Porsch, Anton Aboukhalil, Preston W. Estep, Anastasia Vedenko, Alexandre Palagi, Yongsok Kim, Xianmin Zhu, Brian W. Busser, Caitlin E. Gamble, Antonina Iagovitina, Aditi Singhania, Alan M. Michelson, and Martha L. Bulyk. Highly parallel assays of tissue-specific enhancers in whole *Drosophila* embryos. *Nature Methods*, 10(8):774–780, 2013.

[75] Cosmas D. Arnold, Daniel Gerlach, Daniel Spies, Jessica A. Matts, Yuliya A. Sytnikova, Michaela Pagani, Nelson C. Lau, and Alexander Stark. Quantitative genome-wide enhancer activity maps for five *Drosophila* species show

functional enhancer conservation and turnover during cis-regulatory evolution. *Nature Genetics*, 46(7):685–692, 2014.

[76] Diane E. Dickel, Yiwen Zhu, Alex S. Nord, John N. Wylie, Jennifer A. Akiyama, Veena Afzal, Ingrid Plajzer-Frick, Aileen Kirkpatrick, Berthold Göttgens, Benoit G. Bruneau, Axel Visel, and Len A. Pennacchio. Function-based identification of mammalian enhancers using site-specific integration. *Nature Methods*, 11(5):566–571, 2014.

[77] Wenxue Zhao, Joshua L. Pollack, Denitza P. Blagev, Noah Zaitlen, Michael T. McManus, and David J. Erle. Massively parallel functional annotation of 3' untranslated regions. *Nature Biotechnology*, 32(4):387–391, 2014.

[78] Michael A. White, Connie A. Myers, Joseph C. Corbo, and Barak A. Cohen. Massively parallel *in vivo* enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences*, 110(29):11952–11957, 2013.

[79] Ramon Y. Birnbaum, Rupali P. Patwardhan, Mee J. Kim, Gregory M. Findlay, Beth Martin, Jingjing Zhao, Robert J.A. Bell, Robin P. Smith, Angel A. Ku, Jay Shendure, and Nadav Ahituv. Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation. *PLoS Genetics*, 10(10), 2014.

[80] Michal Levo, Tali Avnit-Sagi, Maya Lotan-Pompan, Yael Kalma, Adina Weinberger, Zohar Yakhini, and Eran Segal. Systematic investigation of transcription factor activity in the context of chromatin using massively parallel binding and expression assays. *Molecular Cell*, 65:604–617, 2017.

[81] Lewis G. Halsey, Douglas Curran-Everett, Sarah L. Vowler, and Gordon B. Drummond. The fickle *P* value generates irreproducible results. *Nature Methods*, 12(3):179–185, 2015.

[82] Devinderjit Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, 2nd edition, 2006.

[83] Michael Lässig. From biophysics to evolutionary genetics: Statistical aspects of gene regulation. *BMC Bioinformatics*, 8 Suppl 6:S7, 2007.

[84] Don van Ravenzwaaij, Pete Cassey, and Scott D. Brown. A simple introduction to Markov Chain Monte Carlo sampling. *Psychonomic Bulletin and Review*, 25:143–154, 2018.

[85] Adrian F. M. Smith and Gareth O. Roberts. Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society*, 55(1):3–23, 1993.

[86] Gurinder S. Atwal and Justin B. Kinney. Learning quantitative sequence–function relationships from massively parallel experiments. *Journal of Statistical Physics*, 162(5):1203–1243, 2016.

Chapter 2

TUNING TRANSCRIPTIONAL REGULATION THROUGH SIGNALING: A PREDICTIVE THEORY OF ALLOSTERIC INDUCTION

A version of this chapter is in press as Manuel Razo-Mejia, Stephanie L. Barnes, Nathan M. Belliveau, Griffin Chure, Tal Einav, Mitchell Lewis, and Rob Phillips. Tuning transcriptional regulation through signaling: A predictive theory of allosteric regulation. *Cell Systems*, In press, 2018.

M.R.M., S.L.B., N.M.B., G.C., T.E. contributed equally to this work.

2.1 Introduction

Understanding how organisms sense and respond to changes in their environment has long been a central theme of biological inquiry. At the cellular level, this interaction is mediated by a diverse collection of molecular signaling pathways. A pervasive mechanism of signaling in these pathways is allosteric regulation, in which the binding of a ligand induces a conformational change in some target molecule, triggering a signaling cascade [1]. One of the most important examples of such signaling is offered by transcriptional regulation, where a transcription factor's propensity to bind to DNA will be altered upon binding to an allosteric effector.

Despite allostery's ubiquity, we lack a formal, rigorous, and generalizable framework for studying its effects across the broad variety of contexts in which it appears. A key example of this is transcriptional regulation, in which allosteric transcription factors can be induced or corepressed by binding to a ligand. An allosteric transcription factor can adopt multiple conformational states, each of which has its own affinity for the ligand and for its DNA target site. *In vitro* studies have rigorously quantified the equilibria of different conformational states for allosteric transcription factors and measured the affinities of these states to the ligand [2, 3]. In spite of these experimental observations, the lack of a coherent quantitative model for allosteric transcriptional regulation has made it impossible to predict the behavior of even a simple genetic circuit across a range of regulatory parameters.

The ability to predict circuit behavior robustly—that is, across both broad ranges of parameters and regulatory architectures—is important for multiple reasons. First, in

the context of a specific gene, accurate prediction demonstrates that all components relevant to the gene's behavior have been identified and characterized to sufficient quantitative precision. Second, in the context of genetic circuits in general, robust prediction validates the model that generated the prediction. Possessing a validated model also has implications for future work. For example, when we have sufficient confidence in the model, a single data set can be used to accurately extrapolate a system's behavior in other conditions. Moreover, there is an essential distinction between a predictive model, which is used to predict a system's behavior given a set of input variables, and a retroactive model, which is used to describe the behavior of data that has already been obtained. We note that even some of the most careful and rigorous analysis of transcriptional regulation often entails only a retroactive reflection on a single experiment. This raises the fear that each regulatory architecture may require a unique analysis that cannot carry over to other systems, a worry that is exacerbated by the prevalent use of phenomenological functions (e.g. Hill functions or ratios of polynomials) that can analyze a single data set but cannot be used to extrapolate a system's behavior in other conditions [4–8].

This work explores what happens when theory takes center stage, namely, we first write down the equations governing a system and describe its expected behavior across a wide array of experimental conditions, and only then do we set out to experimentally confirm these results. Building upon previous work [9–11] and the work of Monod, Wyman, and Changeux [12], we present a statistical mechanical rendering of allostery in the context of induction and corepression (shown schematically in 2.1A and henceforth referred to as the MWC model) and use it as the basis of parameter-free predictions which we then test experimentally. More specifically, we study the simple repression motif—a widespread bacterial genetic regulatory architecture in which binding of a transcription factor occludes binding of an RNA polymerase, thereby inhibiting transcription initiation. The MWC model stipulates that an allosteric protein fluctuates between two distinct conformations—an active and inactive state—in thermodynamic equilibrium [12]. During induction, for example, effector binding increases the probability that a repressor will be in the inactive state, weakening its ability to bind to the promoter and resulting in increased expression. To test the predictions of our model across a wide range of operator binding strengths and repressor copy numbers, we design an *E. coli* genetic construct in which the binding probability of a repressor regulates gene expression of a fluorescent reporter.

In total, the work presented here demonstrates that one extremely compact set of parameters can be applied self-consistently and predictively to different regulatory situations including simple repression on the chromosome, cases in which decoy binding sites for repressor are put on plasmids, cases in which multiple genes compete for the same regulatory machinery, cases involving multiple binding sites for repressor leading to DNA looping, and induction by signaling [9, 10, 13–16]. Thus, rather than viewing the behavior of each circuit as giving rise to its own unique input-output response, the MWC model provides a means to characterize these seemingly diverse behaviors using a single unified framework governed by a small set of parameters.

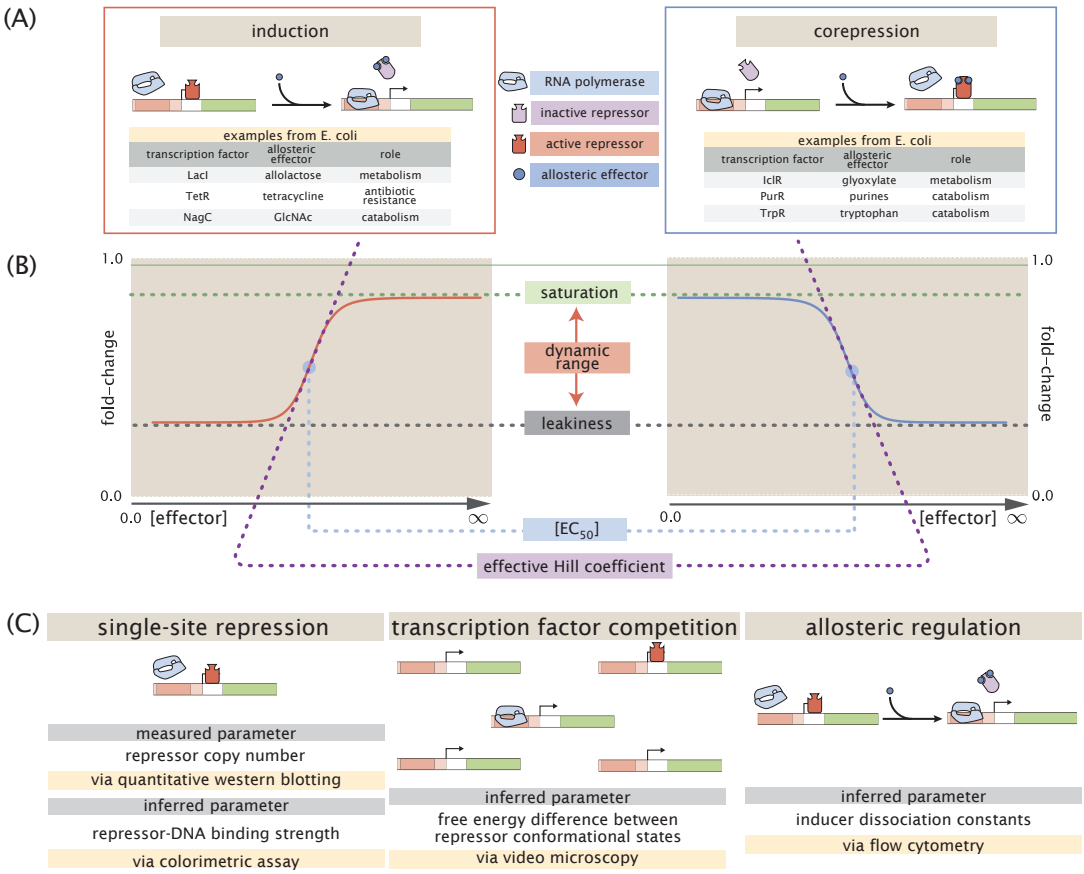

Figure 2.1: **Transcription regulation architectures involving an allosteric repressor.** (A) We consider a promoter regulated solely by an allosteric repressor. When bound, the repressor prevents RNAP from binding and initiating transcription. Induction is characterized by the addition of an effector which binds to the repressor and stabilizes the inactive state (defined as the state which has a low affinity for DNA), thereby increasing gene expression. In corepression, the effector stabilizes the repressor's active state and thus further reduces gene expression. We list several characterized examples of induction and corepression in *E. coli* [17, 18]. (B) A schematic regulatory response of the two architectures shown in Panel A plotting the fold-change in gene expression as a function of effector concentration. Phenotypic properties that describe each response curve include the leakiness, saturation, dynamic range, the concentration of ligand which generates a fold-change halfway between the minimal and maximal response ($[EC_{50}]$), and the log-log slope at the midpoint of the response (effective Hill coefficient). (C) Over time we have refined our understanding of simple repression architectures. A first round of experiments used colorimetric assays and quantitative Western blots to investigate how single-site repression is modified by the repressor copy number and repressor-DNA binding energy [9]. A second round of experiments used video microscopy to probe how the copy number of the promoter and presence of competing repressor binding sites affect gene expression [11]. Here we used flow cytometry to determine the inducer-repressor dissociation constants.

2.2 Results

Characterizing Transcription Factor Induction using the Monod-Wyman-Changeux (MWC) Model

We begin by considering a simple repression genetic architecture in which the binding of an allosteric repressor occludes the binding of RNA polymerase (RNAP) to the DNA [19, 20]. When an effector (hereafter referred to as an "inducer" for the case of induction) binds to the repressor, it shifts the repressor's allosteric equilibrium towards the inactive state as specified by the MWC model [12]. This causes the repressor to bind more weakly to the operator, which increases gene expression. Simple repression motifs in the absence of inducer have been previously characterized by an equilibrium model where the probability of each state of repressor and RNAP promoter occupancy is dictated by the Boltzmann distribution [9, 10, 19–22] (we note that non-equilibrium models of simple repression have been shown to have the same functional form that we derive below [23]). We extend these models to consider allostery by accounting for the equilibrium state of the repressor through the MWC model.

Thermodynamic models of gene expression begin by enumerating all possible states of the promoter and their corresponding statistical weights. As shown in Figure 2.2A, the promoter can either be empty, occupied by RNAP, or occupied by either an active or inactive repressor. The probability of binding to the promoter will be affected by the protein copy number, which we denote as P for RNAP, R_A for active repressor, and R_I for inactive repressor. We note that repressors fluctuate between the active and inactive conformation in thermodynamic equilibrium, such that R_A and R_I will remain constant for a given inducer concentration [12]. We assign the repressor a different DNA binding affinity in the active and inactive state. In addition to the specific binding sites at the promoter, we assume that there are N_{NS} non-specific binding sites elsewhere (i.e. on parts of the genome outside the simple repression architecture) where the RNAP or the repressor can bind. All specific binding energies are measured relative to the average non-specific binding energy. Thus, $\Delta\varepsilon_P$ represents the energy difference between the specific and non-specific binding for RNAP to the DNA. Likewise, $\Delta\varepsilon_{RA}$ and $\Delta\varepsilon_{RI}$ represent the difference in specific and non-specific binding energies for repressor in the active or inactive state, respectively.

Thermodynamic models of transcription [9–11, 19–22, 24–26] posit that gene expression is proportional to the probability that the RNAP is bound to the promoter

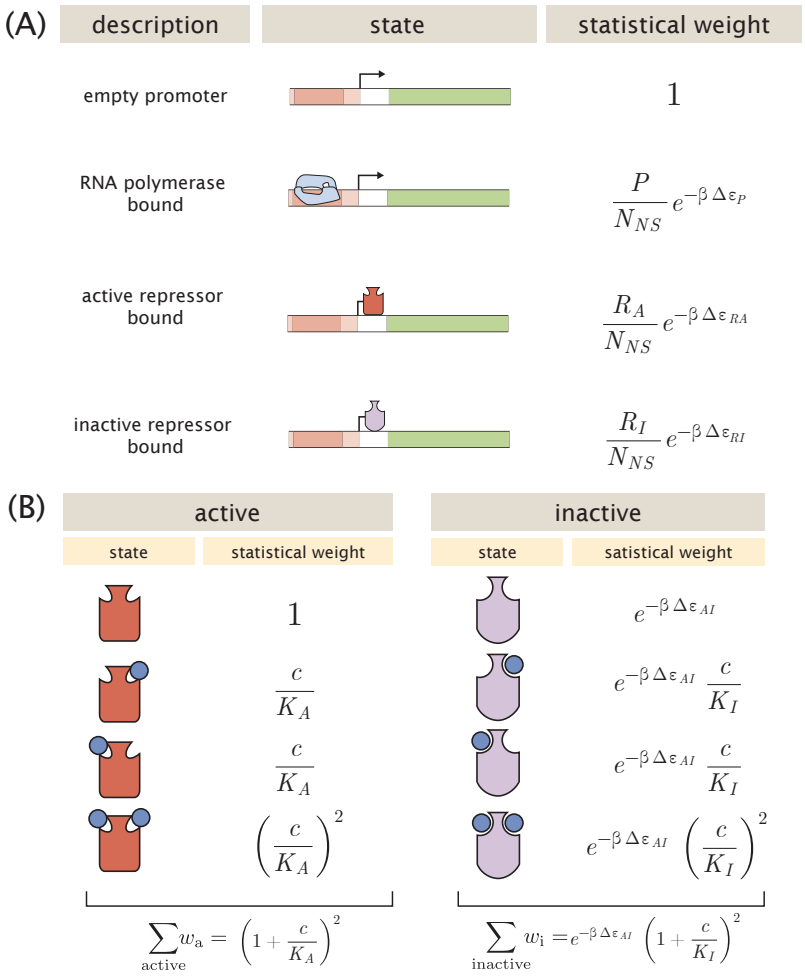

Figure 2.2: **States and weights for the simple repression motif.** (A) RNAP (light blue) and a repressor compete for binding to a promoter of interest. There are R_A repressors in the active state (red) and R_I repressors in the inactive state (purple). The difference in energy between a repressor bound to the promoter of interest versus another non-specific site elsewhere on the DNA equals $\Delta\varepsilon_{RA}$ in the active state and $\Delta\varepsilon_{RI}$ in the inactive state; the P RNAP have a corresponding energy difference $\Delta\varepsilon_P$ relative to non-specific binding on the DNA. N_{NS} represents the number of non-specific binding sites for both RNAP and repressor. (B) A repressor has an active conformation (red, left column) and an inactive conformation (purple, right column), with the energy difference between these two states given by $\Delta\varepsilon_{AI}$. The inducer (blue circle) at concentration c is capable of binding to the repressor with dissociation constants K_A in the active state and K_I in the inactive state. The eight states for a dimer with $n = 2$ inducer binding sites are shown along with the sums of the active and inactive states.

p_{bound}, which is given by

$$p_{\text{bound}} = \frac{\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}{1 + \frac{R_A}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} + \frac{R_I}{N_{NS}} e^{-\beta\Delta\varepsilon_{RI}} + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}, \tag{2.1}$$

with $\beta = \frac{1}{k_B T}$ where k_B is the Boltzmann constant and T is the temperature of the system. As $k_B T$ is the natural unit of energy at the molecular length scale, we treat the products $\beta\Delta\varepsilon_j$ as single parameters within our model. Measuring p_{bound} directly is fraught with experimental difficulties, as determining the exact proportionality between expression and p_{bound} is not straightforward. Instead, we measure the fold-change in gene expression due to the presence of the repressor. We define fold-change as the ratio of gene expression in the presence of repressor relative to expression in the absence of repressor (i.e. constitutive expression), namely,

$$\text{fold-change} \equiv \frac{p_{\text{bound}}(R > 0)}{p_{\text{bound}}(R = 0)}. \tag{2.2}$$

We can simplify this expression using two well-justified approximations: (1) $\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P} \ll 1$ implying that the RNAP binds weakly to the promoter ($N_{NS} = 4.6 \times 10^6$, $P \approx 10^3$ [27], $\Delta\varepsilon_P \approx -2$ to $-5\ k_B T$ [14], so that $\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P} \approx 0.01$) and (2) $\frac{R_I}{N_{NS}} e^{-\beta\Delta\varepsilon_{RI}} \ll 1 + \frac{R_A}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}$ which reflects our assumption that the inactive repressor binds weakly to the promoter of interest. Using these approximations, the fold-change reduces to the form

$$\text{fold-change} \approx \left(1 + \frac{R_A}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1} \equiv \left(1 + p_A(c) \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{2.3}$$

where in the last step we have introduced the fraction $p_A(c)$ of repressors in the active state given a concentration c of inducer, such that $R_A(c) = p_A(c)R$. Since inducer binding shifts the repressors from the active to the inactive state, $p_A(c)$ grows smaller as c increases [28].

We use the MWC model to compute the probability $p_A(c)$ that a repressor with n inducer binding sites will be active. The value of $p_A(c)$ is given by the sum of the weights of the active repressor states divided by the sum of the weights of all possible repressor states (see Figure 2.2B), namely,

$$p_A(c) = \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta\Delta\varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n}, \tag{2.4}$$

where K_A and K_I represent the dissociation constant between the inducer and repressor in the active and inactive states, respectively, and $\Delta\varepsilon_{AI} = \varepsilon_I - \varepsilon_A$ is

the free energy difference between a repressor in the inactive and active state (the quantity $e^{-\Delta\varepsilon_{AI}}$ is sometimes denoted by L [12, 28] or $K_{\mathrm{RR}*}$ [26]). In this equation, $\frac{c}{K_A}$ and $\frac{c}{K_I}$ represent the change in free energy when an inducer binds to a repressor in the active or inactive state, respectively, while $e^{-\beta\Delta\varepsilon_{AI}}$ represents the change in free energy when the repressor changes from the active to inactive state in the absence of inducer. Thus, a repressor which favors the active state in the absence of inducer ($\Delta\varepsilon_{AI} > 0$) will be driven towards the inactive state upon inducer binding when $K_I < K_A$. The specific case of a repressor dimer with $n = 2$ inducer binding sites is shown in Figure 2.2B.

Substituting $p_A(c)$ from Equation 2.4 into Equation 2.3 yields the general formula for induction of a simple repression regulatory architecture [23], namely,

$$\text{fold-change} = \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta\Delta\varepsilon_{AI}}\left(1 + \frac{c}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}. \tag{2.5}$$

While we have used the specific case of simple repression with induction to craft this model, the same mathematics describe the case of corepression in which binding of an allosteric effector stabilizes the active state of the repressor and decreases gene expression (see Figure 2.1B). Interestingly, we shift from induction (governed by $K_I < K_A$) to corepression ($K_I > K_A$) as the ligand transitions from preferentially binding to the inactive repressor state to stabilizing the active state. Furthermore, this general approach can be used to describe a variety of other motifs such as activation, multiple repressor binding sites, and combinations of activator and repressor binding sites [10, 11, 24].

The formula presented in Equation 2.5 enables us to make precise quantitative statements about induction profiles. Motivated by the broad range of predictions implied by Equation 2.5, we designed a series of experiments using the *lac* system in *E. coli* to tune the control parameters for a simple repression genetic circuit. As discussed in Figure 2.1C, previous studies from our lab have provided well-characterized values for many of the parameters in our experimental system, leaving only the values of the the MWC parameters (K_A, K_I, and $\Delta\varepsilon_{AI}$) to be determined. We note that while previous studies have obtained values for K_A, K_I, and $L = e^{-\beta\Delta\varepsilon_{AI}}$ [26, 29], they were either based upon biochemical experiments or *in vivo* conditions involving poorly characterized transcription factor copy numbers and gene copy numbers. These differences relative to our experimental conditions and fitting techniques led us to believe that it was important to perform our own analysis of

these parameters. After inferring these three MWC parameters (see Supplemental Section 2.5 for details regarding the inference of $\Delta\varepsilon_{AI}$, which was fitted separately from K_A and K_I), we were able to predict the input/output response of the system under a broad range of experimental conditions. For example, this framework can predict the response of the system at different repressor copy numbers R, repressor-operator affinities $\Delta\varepsilon_{RA}$, inducer concentrations c, and gene copy numbers (see Supplemental Section 2.6).

Experimental Design

We test our model by predicting the induction profiles for an array of strains that could be made using previously characterized repressor copy numbers and DNA binding energies. Our approach contrasts with previous studies that have parameterized induction curves of simple repression motifs, as these have relied on expression systems where proteins are expressed from plasmids, resulting in highly variable and unconstrained copy numbers [26, 30–33]. Instead, our approach relies on a foundation of previous work as depicted in Figure 2.1C. This includes work from our laboratory that used *E. coli* constructs based on components of the *lac* system to demonstrate how the *lac* repressor (LacI) copy number R and operator binding energy $\Delta\varepsilon_{RA}$ affect gene expression in the absence of inducer [9]. Ref. [34] extended the theory used in that work to the case of multiple promoters competing for a given transcription factor, which was validated experimentally by Ref. [10], who modified this system to consider expression from multiple-copy plasmids as well as the presence of competing repressor binding sites.

The present study extends this body of work by introducing three additional biophysical parameters—$\Delta\varepsilon_{AI}$, K_A, and K_I—which capture the allosteric nature of the transcription factor and complement the results shown by Ref. [9] and Ref. [10]. Although the current work focuses on systems with a single site of repression, in Section 2.5 we utilize data from Ref. [10], in which multiple sites of repression are explored, to characterize the allosteric free energy difference $\Delta\varepsilon_{AI}$ between the repressor's active and inactive states. As explained in that Section, this additional data set is critical because multiple degenerate sets of parameters can characterize an induction curve equally well, with the $\Delta\varepsilon_{AI}$ parameter compensated by the inducer dissociation constants K_A and K_I (see Figure 2.8). After fixing $\Delta\varepsilon_{AI}$ as described in the Section 2.5, we can use data from single-site simple repression systems to determine the values of K_A and K_I.

We determine the values of K_A and K_I by fitting to a single induction profile using Bayesian inferential methods [35]. We then use Equation 2.5 to predict gene expression for any concentration of inducer, repressor copy number, and DNA binding energy and compare these predictions against experimental measurements. To obtain induction profiles for a set of strains with varying repressor copy numbers, we used modified *lacI* ribosomal binding sites from Ref. [9] to generate strains with mean repressor copy number per cell of $R = 22 \pm 4$, 60 ± 20, 124 ± 30, 260 ± 40, 1220 ± 160, and 1740 ± 340, where the error denotes standard deviation

of at least three replicates as measured by Ref. [9]. We note that R refers to the number of repressor dimers in the cell, which is twice the number of repressor tetramers reported by Ref. [9]; since both heads of the repressor are assumed to always be either specifically or non-specifically bound to the genome, the two repressor dimers in each LacI tetramer can be considered independently. Gene expression was measured using a Yellow Fluorescent Protein (YFP) gene, driven by a *lacUV5* promoter. Each of the six repressor copy number variants were paired with the native O1, O2, or O3 *lac* operator [36] placed at the YFP transcription start site, thereby generating eighteen unique strains. The repressor-operator binding energies (O1 $\Delta\varepsilon_{RA} = -15.3 \pm 0.2\ k_BT$, O2 $\Delta\varepsilon_{RA} = -13.9\ k_BT \pm 0.2$, and O3 $\Delta\varepsilon_{RA} = -9.7 \pm 0.1\ k_BT$) were previously inferred by measuring the fold-change of the *lac* system at different repressor copy numbers, where the error arises from model fitting [9]. Additionally, we were able to obtain the value $\Delta\varepsilon_{AI} = 4.5\ k_BT$ by fitting to previous data as discussed in Section 2.5. We measure fold-change over a range of known IPTG concentrations c, using $n = 2$ inducer binding sites per LacI dimer and approximating the number of non-specific binding sites as the length in base-pairs of the *E. coli* genome, $N_{NS} = 4.6 \times 10^6$.

Our experimental pipeline for determining fold-change using flow cytometry is shown in Figure 2.3. Briefly, cells were grown to exponential phase, in which gene expression reaches steady state [37], under concentrations of the inducer IPTG ranging between 0 and 5 mM. We measure YFP fluorescence using flow cytometry and automatically gate the data to include only single-cell measurements (see Section 2.7). To validate the use of flow cytometry, we also measured the fold-change of a subset of strains using the established method of single-cell microscopy (see Supplemental Section 2.8). We found that the fold-change measurements obtained from microscopy were indistinguishable from that of flow-cytometry and yielded values for the inducer binding constants K_A and K_I that were within error.

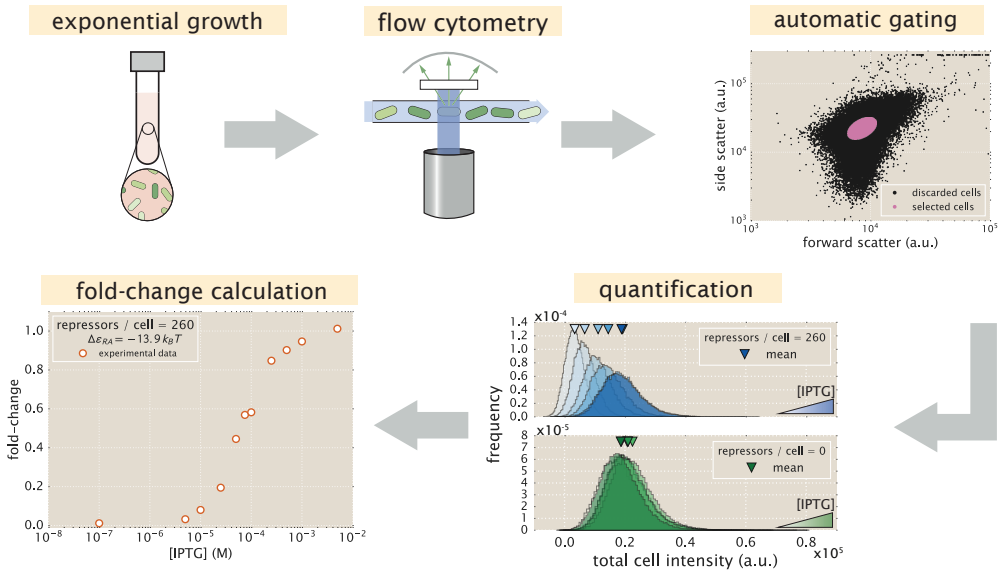

Figure 2.3: **An experimental pipeline for high-throughput fold-change measurements.** Cells are grown to exponential steady state and their fluorescence is measured using flow cytometry. Automatic gating methods using forward- and side-scattering are used to ensure that all measurements come from single cells (see Methods). Mean expression is then quantified at different IPTG concentrations (top, blue histograms) and for a strain without repressor (bottom, green histograms), which shows no response to IPTG as expected. Fold-change is computed by dividing the mean fluorescence in the presence of repressor by the mean fluorescence in the absence of repressor.

Determination of the *in vivo* MWC Parameters

The three parameters that we tune experimentally are shown in Figure 2.4A, leaving the three allosteric parameters ($\Delta\varepsilon_{AI}$, K_A, and K_I) to be determined by fitting. We used previous LacI fold-change data [10] to infer that $\Delta\varepsilon_{AI} = 4.5\ k_BT$ (see Supplemental Section 2.5). Rather than fitting K_A and K_I to our entire data set of eighteen unique constructs, we performed Bayesian parameter estimation on data from a single strain with $R = 260$ and an O2 operator ($\Delta\varepsilon_{RA} = -13.9\ k_BT$ [9]) shown in Figure 2.4D (white circles). Using Markov Chain Monte Carlo, we determine the most likely parameter values to be $K_A = 139^{+29}_{-22} \times 10^{-6}$ M and $K_I = 0.53^{+0.04}_{-0.04} \times 10^{-6}$ M, which are the modes of their respective distributions, where the superscripts and subscripts represent the upper and lower bounds of the 95$^{\text{th}}$ percentile of the parameter value distributions (see Figure 2.4B). Unfortunately, we are not able to make a meaningful value-for-value comparison of our parameters to those of earlier studies [26, 31] because of uncertainties in both gene copy number and transcription factor copy numbers in these studies, as illustrated by the plots in Section 2.6. We then predicted the fold-change for the remaining seventeen strains with no further fitting (see Figure 2.4C-E) together with the specific phenotypic properties described in Figure 2.1 and discussed in detail below (see Figure 2.4F-J). The shaded regions in Figure 2.4C-J denote the 95% credible regions. Factors determining the width of the credible regions are explored in Supplemental Section 2.9.

We stress that the entire suite of predictions in Figure 2.4 is based upon the induction profile of a single strain. Our ability to make such a broad range of predictions stems from the fact that our parameters of interest—such as the repressor copy number and DNA binding energy—appear as distinct physical parameters within our model. While the single data set in Figure 2.4D could also be fit using a Hill function, such an analysis would be unable to predict any of the other curves in the figure (see Supplemental Section 2.10). Phenomenological expressions such as the Hill function can describe data, but lack predictive power and are thus unable to build our intuition, help us design *de novo* input-output functions, or guide future experiments [25, 30].

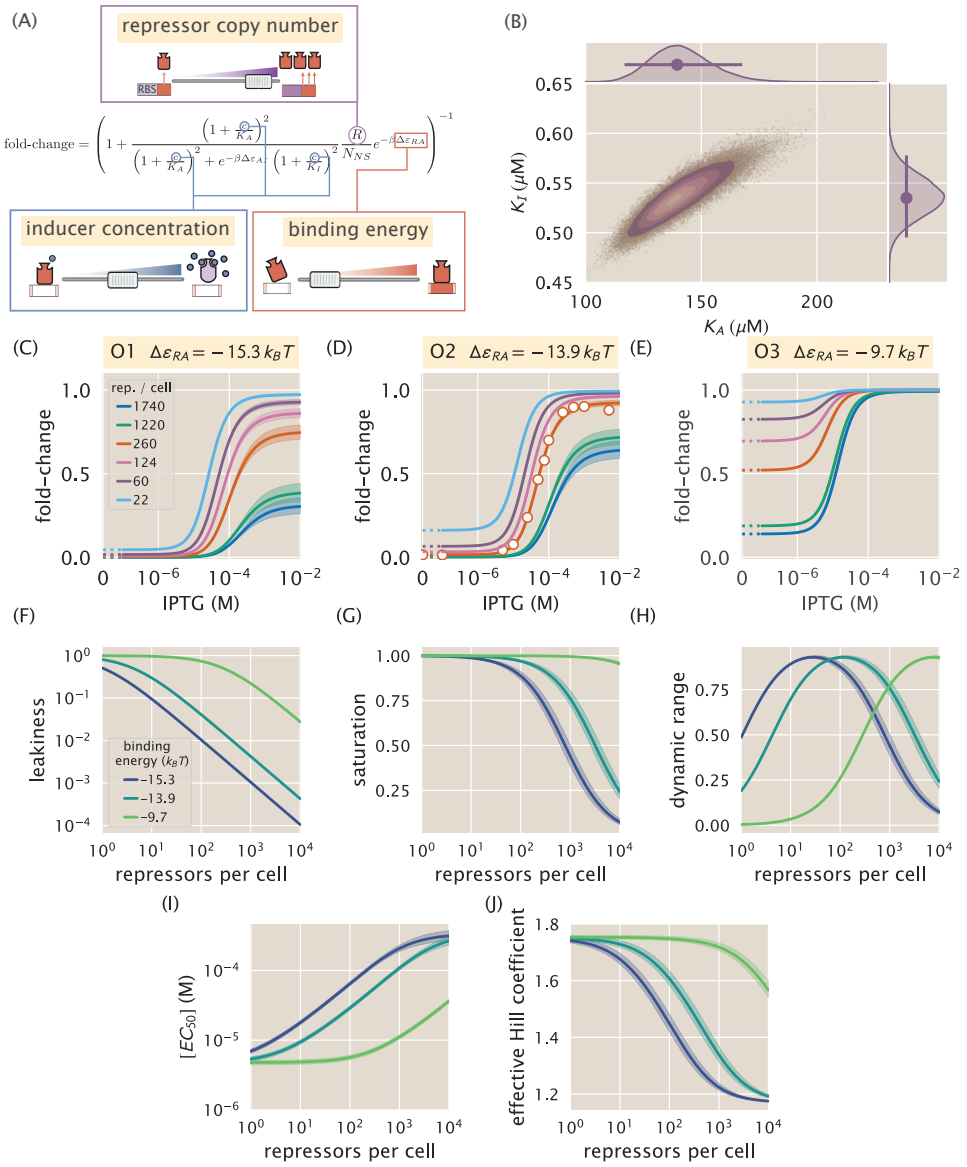

Figure 2.4: **Predicting induction profiles for different biological control parameters.** (A) We can quantitatively tune R via ribosomal binding site (RBS) modifications, $\Delta\varepsilon_{RA}$ by mutating the operator sequence, and c by adding different amounts of IPTG to the growth medium. (B) The unknown dissociation constants K_A and K_I between the inducer and the repressor in the active and inactive states, respectively, can be inferred using Bayesian parameter estimation from a single induction curve. (C-E) Predicted IPTG titration curves for different repressor copy numbers and operator strengths. Values for K_A and K_I are fitted to titration data for the O2 strain (white circles in Panel D) with $R = 260$, $\Delta\varepsilon_{RA} = -13.9\ k_BT$, $n = 2$, and $\Delta\varepsilon_{AI} = 4.5\ k_BT$. The remaining solid lines predict the fold-change Equation 2.5 all other strains. Error bars of experimental data show the standard error of the mean (eight or more replicates) when this error is not smaller than the diameter of the data point. The shaded regions denote the 95% credible region, although the credible region is obscured when it is thinner than the curve itself. Additionally, our model allows us to investigate key phenotypic properties of the induction profiles (see Figure 2.1B). Specifically, we show predictions for the (F) leakiness, (G) saturation, (H) dynamic range,(I) $[EC_{50}]$, and (J) effective Hill coefficient of the induction profiles.

Comparison of Experimental Measurements with Theoretical Predictions

We tested the predictions shown in Figure 2.4 by measuring fold-change induction profiles in strains with a broad range of repressor copy numbers and repressor binding energies as characterized in Ref. [9]. With a few notable exceptions, the results shown in Figure 2.5 demonstrate agreement between theory and experiment. We note that there was an apparently systematic shift in the O3 $\Delta\varepsilon_{RA} = -9.7\ k_BT$ strains (Figure 2.5C) and all of the $R = 1220$ and $R = 1740$ strains. This may be partially due to imprecise previous determinations of their $\Delta\varepsilon_{RA}$ and R values. By performing a global fit where we infer all parameters including the repressor copy number R and the binding energy $\Delta\varepsilon_{RA}$, we found better agreement for these strains, although a discrepancy in the steepness of the response for all O3 strains remains (see Supplemental Section 2.11). We considered a number of hypotheses to explain these discrepancies such as including other states (e.g. non-negligible binding of the inactive repressor), relaxing the weak promoter approximation, and accounting for variations in gene and repressor copy number throughout the cell cycle, but none explained the observed discrepancies. As an additional test of our model, we considered strains using the synthetic Oid operator which exhibits an especially strong binding energy of $\Delta\varepsilon_{RA} = -17\ k_BT$ [9]. The global fit agrees well with the Oid microscopy data, though it asserts a stronger Oid binding energy of $\Delta\varepsilon_{RA} = -17.7\ k_BT$ (see Supplemental Section 2.12).

To ensure that the agreement between our predictions and data is not an accident of the strain we used to perform our fitting, we also inferred K_A and K_I from each of the other strains. As shown in Supplemental Section 2.13 and Figure 2.5D, the inferred values of K_A and K_I depend minimally upon which strain is chosen, indicating that these parameter values are highly robust. We also performed a global fit using the data from all eighteen strains in which we fitted for the inducer dissociation constants K_A and K_I, the repressor copy number R, and the repressor DNA binding energy $\Delta\varepsilon_{RA}$ (see Supplemental Section 2.11). The resulting parameter values were nearly identical to those fitted from any single strain. For the remainder of the text we continue using parameters fitted from the strain with $R = 260$ repressors and an O2 operator.

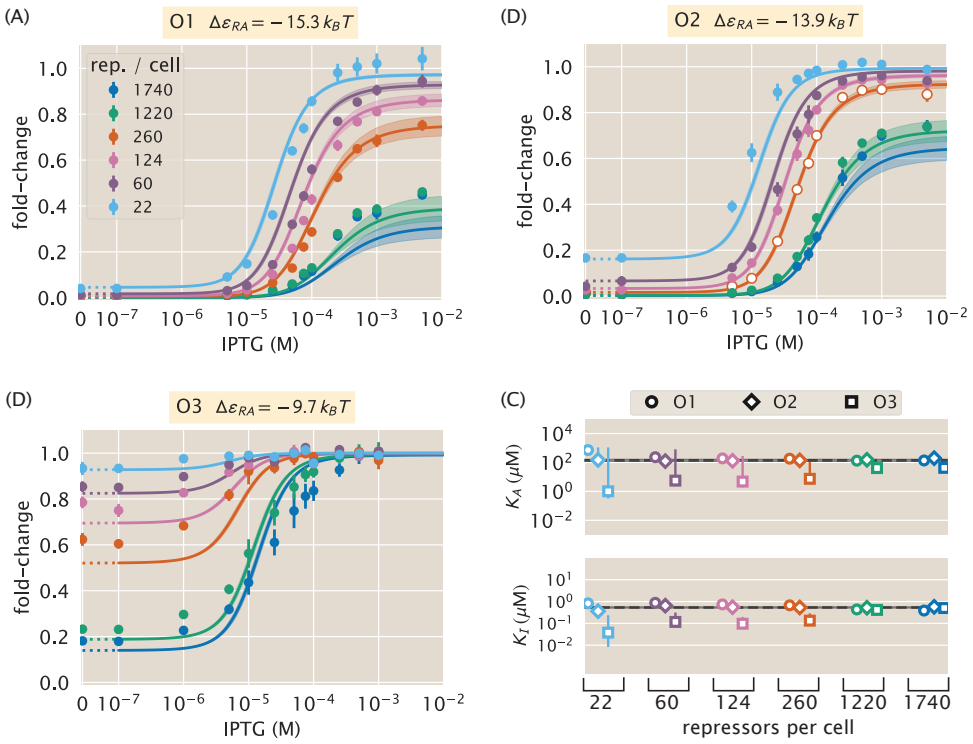

Figure 2.5: **Comparison of predictions against measured and inferred data.** Flow cytometry measurements of fold-change over a range of IPTG concentrations for (A) O1, (B) O2, and (C) O3 strains at varying repressor copy numbers, overlaid on the predicted responses. Error bars for the experimental data show the standard error of the mean (eight or more replicates). As discussed in Figure 2.4, all of the predicted induction curves were generated prior to measurement by inferring the MWC parameters using a single data set (O2 $R = 260$, shown by white circles in Panel B). The predictions may therefore depend upon which strain is used to infer the parameters. (D) The inferred parameter values of the dissociation constants K_A and K_I using any of the eighteen strains instead of the O2 $R = 260$ strain. Nearly identical parameter values are inferred from each strain, demonstrating that the same set of induction profiles would have been predicted regardless of which strain was chosen. The points show the mode, and the error bars denote the 95% credible region of the parameter value distribution. Error bars not visible are smaller than the size of the marker.

Predicting the Phenotypic Traits of the Induction Response

A subset of the properties shown in Figure 2.1 (i.e. the leakiness, saturation, dynamic range, $[EC_{50}]$, and effective Hill coefficient) are of significant interest to synthetic biology. For example, synthetic biology is often focused on generating large responses (i.e. a large dynamic range) or finding a strong binding partner (i.e. a small $[EC_{50}]$) [38, 39]. While these properties are all individually informative, when taken together they capture the essential features of the induction response. We reiterate that a Hill function approach cannot predict these features *a priori* and furthermore requires fitting each curve individually. The MWC model, on the other hand, enables us to quantify how each trait depends upon a single set of physical parameters as shown by Figure 2.4F-J.

We define these five phenotypic traits using expressions derived from the model, Equation 2.5. These results build upon extensive work by Ref. [40], who computed many such properties for ligand-receptor binding within the MWC model. We begin by analyzing the leakiness, which is the minimum fold-change observed in the absence of ligand, given by

$$\begin{aligned}\text{leakiness} &= \text{fold-change}(c = 0) \\ &= \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1},\end{aligned} \tag{2.6}$$

and the saturation, which is the maximum fold change observed in the presence of saturating ligand,

$$\begin{aligned}\text{saturation} &= \text{fold-change}(c \to \infty) \\ &= \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}.\end{aligned} \tag{2.7}$$

Systems that minimize leakiness repress strongly in the absence of effector while systems that maximize saturation have high expression in the presence of effector. Together, these two properties determine the dynamic range of a system's response, which is given by the difference

$$\text{dynamic range} = \text{saturation} - \text{leakiness}. \tag{2.8}$$

These three properties are shown in Figure 2.4F-H. We discuss these properties in greater detail in Supplemental Section 2.14. Figure 2.6A-C shows that the

measurements of these three properties, derived from the fold-change data in the absence of IPTG and the presence of saturating IPTG, closely match the predictions for all three operators.

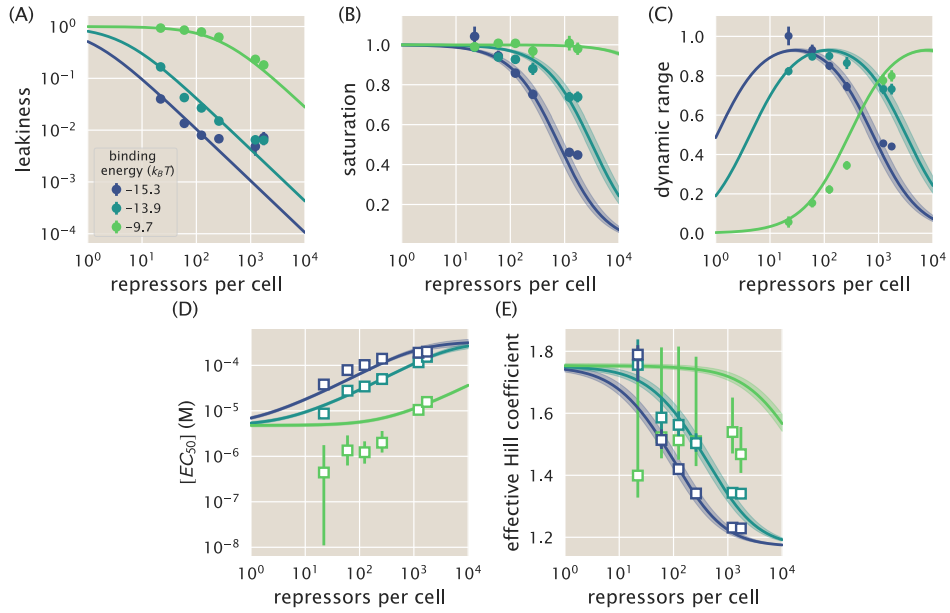

Figure 2.6: **Predictions and experimental measurements of key properties of induction profiles.** Data for the (A) leakiness, (B) saturation, and (C) dynamic range are obtained from fold-change measurements in Figure 2.5 in the absence of IPTG and at saturating concentrations of IPTG. The three repressor-operator binding energies in the legend correspond to the O1 operator (−15.3 k_BT), O2 operator (−13.9 k_BT), and O3 operator (−9.7 k_BT). Both the (D) [EC_{50}] and (D) effective Hill coefficient are inferred by individually fitting each operator-repressor pairing in Figure 2.5A-C separately to Equation 2.5 in order to smoothly interpolate between the data points. Error bars for A-C represent the standard error of the mean for eight or more replicates; error bars for D-E represent the 95% credible region for the parameter found by propagating the credible region of our estimates of K_A and K_I into Equation 2.9 and Equation 2.10.

Two additional properties of induction profiles are the [EC_{50}] and effective Hill coefficient, which determine the range of inducer concentration in which the system's output goes from its minimum to maximum value. The [EC_{50}] denotes the inducer concentration required to generate a system response Equation 2.5 halfway between its minimum and maximum value,

$$\text{fold-change}(c = [EC_{50}]) = \frac{\text{leakiness} + \text{saturation}}{2}. \tag{2.9}$$

The effective Hill coefficient h, which quantifies the steepness of the curve at the

$[EC_{50}]$ [28], is given by

$$h = \left(2\frac{d}{d\log c}\left[\log\left(\frac{\text{fold-change}(c) - \text{leakiness}}{\text{dynamic range}}\right)\right]\right)_{c=[EC_{50}]}. \tag{2.10}$$

Figure 2.4I-J shows how the $[EC_{50}]$ and effective Hill coefficient depend on the repressor copy number. In Supplemental Section 2.14, we discuss the analytic forms of these two properties as well as their dependence on the repressor-DNA binding energy.

Figure 2.6D-E shows the estimated values of the $[EC_{50}]$ and the effective Hill coefficient overlaid on the theoretical predictions. Both properties were obtained by fitting Equation 2.5 to each individual titration curve and computing the $[EC_{50}]$ and effective Hill coefficient using Equation 2.9 and Equation 2.10, respectively. We find that the predictions made with the single strain fit closely match those made for each of the strains with O1 and O2 operators, but the predictions for the O3 operator are markedly off. In Supplemental Section 2.10, we show that the large, asymmetric error bars for the O3 $R = 22$ strain arise from its nearly flat response, where the lack of dynamic range makes it impossible to determine the value of the inducer dissociation constants K_A and K_I, as can be seen in the uncertainty of both the $[EC_{50}]$ and effective Hill coefficient. Discrepancies between theory and data for O3 are improved, but not fully resolved, by performing a global fit or fitting the MWC model individually to each curve (see Supplemental Sections 2.11 and 2.13). It remains an open question how to account for discrepancies in O3, in particular regarding the significant mismatch between the predicted and fitted effective Hill coefficients.

Data Collapse of Induction Profiles

Our primary interest heretofore was to determine the system response at a specific inducer concentration, repressor copy number, and repressor-DNA binding energy. However, the cell does not necessarily "care about" the precise number of repressors in the system or the binding energy of an individual operator. The relevant quantity for cellular function is the fold-change enacted by the regulatory system. This raises the question: given a specific value of the fold-change, what combination of parameters will give rise to this desired response? In other words, what trade-offs between the parameters of the system will give rise to the same mean cellular output? These are key questions both for understanding how the system is governed and for engineering specific responses in a synthetic biology context. To address these questions, we follow the data collapse strategy used in a number of previous studies [41–43], and rewrite Equation 2.5 as a Fermi function,

$$\text{fold-change} = \frac{1}{1 + e^{-F(c)}}, \tag{2.11}$$

where $F(c)$ is the free energy of the repressor binding to the operator of interest relative to the unbound operator state in k_BT units [23, 42, 43], which is given by

$$F(c) = \frac{\Delta\varepsilon_{RA}}{k_BT} - \log \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta\Delta\varepsilon_{AI}}\left(1 + \frac{c}{K_I}\right)^n} - \log \frac{R}{N_{NS}}. \tag{2.12}$$

The first term in $F(c)$ denotes the repressor-operator binding energy, the second the contribution from the inducer concentration, and the last the effect of the repressor copy number. We note that elsewhere, this free energy has been dubbed the Bohr parameter since such families of curves are analogous to the shifts in hemoglobin binding curves at different pHs known as the Bohr effect [23, 44, 45].

Instead of analyzing each induction curve individually, the free energy provides a natural means to simultaneously characterize the diversity in our eighteen induction profiles. Figure 2.7A demonstrates how the various induction curves from Figure 2.4C-E all collapse onto a single master curve, where points from every induction profile that yield the same fold-change are mapped onto the same free energy. Figure 2.7B shows this data collapse for the 216 data points in Figure 2.5A-C, demonstrating the close match between the theoretical predictions and experimental measurements across all eighteen strains.

There are many different combinations of parameter values that can result in the same free energy as defined in Equation 2.12. For example, suppose a system

originally has a fold-change of 0.2 at a specific inducer concentration, and then operator mutations increase the $\Delta\varepsilon_{RA}$ binding energy [46]. While this serves to initially increase both the free energy and the fold-change, a subsequent increase in the repressor copy number could bring the cell back to the original fold-change level. Such trade-offs hint that there need not be a single set of parameters that evoke a specific cellular response, but rather that the cell explores a large but degenerate space of parameters with multiple, equally valid paths.

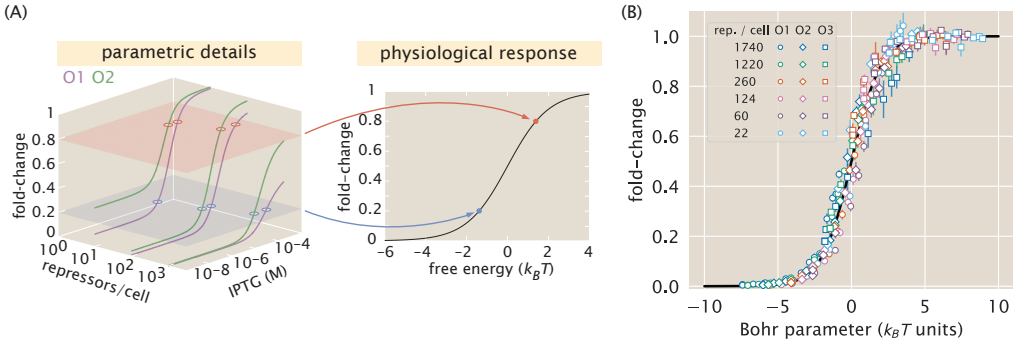

Figure 2.7: **Fold-change data from a broad collection of different strains collapse onto a single master curve.** (A) Any combination of parameters can be mapped to a single physiological response (i.e. fold-change) via the free energy, which encompasses the parametric details of the model. (B) Experimental data from Figure 2.5 collapse onto a single master curve as a function of the free energy Equation 2.12. The free energy for each strain was calculated from Equation 2.12 using $n = 2$, $\Delta\varepsilon_{AI} = 4.5\ k_BT$, $K_A = 139 \times 10^{-6}$ M, $K_I = 0.53 \times 10^{-6}$ M, and the strain-specific R and $\Delta\varepsilon_{RA}$. All data points represent the mean, and error bars are the standard error of the mean for eight or more replicates.

2.3 Discussion

Since the early work by Monod, Wyman, and Changeux [12, 47], an array of biological phenomena have been tied to the existence of macromolecules that switch between inactive and active states. Examples can be found in a wide variety of cellular processes, including ligand-gated ion channels [48], enzymatic reactions [45, 49], chemotaxis [42], quorum sensing [43], G-protein coupled receptors [50], physiologically important proteins [51, 52], and beyond. One of the most ubiquitous examples of allostery is in the context of gene expression, where an array of molecular players bind to transcription factors to influence their ability to regulate gene activity [17, 18]. A number of studies have focused on developing a quantitative understanding of allosteric regulatory systems. Ref. [28, 40] analytically derived fundamental properties of the MWC model, including the leakiness and dynamic range described in this work, noting the inherent trade-offs in these properties when tuning the model's parameters. Work in the Church and Voigt labs, among others, has expanded on the availability of allosteric circuits for synthetic biology [7, 8, 53, 54]. Recently, Daber *et al.* theoretically explored the induction of simple repression within the MWC model [31] and experimentally measured how mutations alter the induction profiles of transcription factors [26]. Vilar and Saiz analyzed a variety of interactions in inducible *lac*-based systems including the effects of oligomerization and DNA folding on transcription factor induction [6, 55]. Other work has attempted to use the *lac* system to reconcile *in vitro* and *in vivo* measurements [33, 56].

Although this body of work has done much to improve our understanding of allosteric transcription factors, there have been few attempts to explicitly connect quantitative models to experiments. Here, we generate a predictive model of allosteric transcriptional regulation and then test the model against a thorough set of experiments using well-characterized regulatory components. Specifically, we used the MWC model to build upon a well-established thermodynamic model of transcriptional regulation [9, 24], allowing us to compose the model from a minimal set of biologically meaningful parameters. This model combines both theoretical and experimental insights; for example, rather than considering gene expression directly we analyze the fold-change in expression, where the weak promoter approximation (see Equation 2.3) circumvents uncertainty in the RNAP copy number. The resulting model depended upon experimentally accessible parameters, namely, the repressor copy number, the repressor-DNA binding energy, and the concentration of inducer. We tested these predictions on a range of strains whose repressor copy

number spanned two orders of magnitude and whose DNA binding affinity spanned 6 k_BT. We argue that one would not be able to generate such a wide array of predictions by using a Hill function, which abstracts away the biophysical meaning of the parameters into phenomenological parameters [57].

More precisely, we tested our model in the context of a *lac*-based simple repression system by first determining the allosteric dissociation constants K_A and K_I from a single induction data set (O2 operator with binding energy $\Delta\varepsilon_{RA} = -13.9\ k_BT$ and repressor copy number $R = 260$) and then using these values to make parameter-free predictions of the induction profiles for seventeen other strains where $\Delta\varepsilon_{RA}$ and R were varied significantly (see Figure 2.4). We next measured the induction profiles of these seventeen strains using flow cytometry and found that our predictions consistently and accurately captured the primary features for each induction data set, as shown in Figure 2.5A-C. Importantly, we find that fitting K_A and K_I to data from any other strain would have resulted in nearly identical predictions (see Figure 2.5D and Supplemental Section 2.13). This suggests that a few carefully chosen measurements can lead to a deep quantitative understanding of how simple regulatory systems work without requiring an extensive sampling of strains that span the parameter space. Moreover, the fact that we could consistently achieve reliable predictions after fitting only two free parameters stands in contrast to the common practice of fitting several free parameters simultaneously, which can nearly guarantee an acceptable fit provided that the model roughly resembles the system response, regardless of whether the details of the model are tied to any underlying molecular mechanism.

Beyond observing changes in fold-change as a function of effector concentration, our application of the MWC model allows us to explicitly predict the values of the induction curves' key parameters, namely, the leakiness, saturation, dynamic range, $[EC_{50}]$, and the effective Hill coefficient (see Figure 2.6). We are consistently able to accurately predict the leakiness, saturation, and dynamic range for each of the strains. For both the O1 and O2 data sets, our model also accurately predicts the effective Hill coefficient and $[EC_{50}]$, though these predictions for O3 are noticeably less accurate. While performing a global fit for all model parameters marginally improves the prediction for O3 (see Supplemental Section 2.11), we are still unable to accurately predict the effective Hill coefficient or the $[EC_{50}]$. We further tried including additional states (such as allowing the inactive repressor to bind to the operator), relaxing the weak promoter approximation, accounting for changes in

gene and repressor copy number throughout the cell cycle [58], and refitting the original binding energies from Ref. [13], but we were still unable to account for the O3 data. It remains an open question as to how the discrepancy between the theory and measurements for O3 can be reconciled.

The dynamic range, which is of considerable interest when designing or characterizing a genetic circuit, is revealed to have an interesting property: although changing the value of $\Delta\varepsilon_{RA}$ causes the dynamic range curves to shift to the right or left, each curve has the same shape and in particular the same maximum value. This means that strains with strong or weak binding energies can attain the same dynamic range when the value of R is tuned to compensate for the binding energy. This feature is not immediately apparent from the IPTG induction curves, which show very low dynamic ranges for several of the O1 and O3 strains. Without the benefit of models that can predict such phenotypic traits, efforts to engineer genetic circuits with allosteric transcription factors must rely on trial and error to achieve specific responses [7, 8].

Despite the diversity observed in the induction profiles of each of our strains, our data are unified by their reliance on fundamental biophysical parameters. In particular, we have shown that our model for fold-change can be rewritten in terms of the free energy Equation 2.12, which encompasses all of the physical parameters of the system. This has proven to be an illuminating technique in a number of studies of allosteric proteins [41–43]. Although it is experimentally straightforward to observe system responses to changes in effector concentration c, framing the input-output function in terms of c can give the misleading impression that changes in system parameters lead to fundamentally altered system responses. Alternatively, if one can find the "natural variable" that enables the output to collapse onto a single curve, it becomes clear that the system's output is not governed by individual system parameters, but rather the contributions of multiple parameters that define the natural variable. When our fold-change data are plotted against the respective free energies for each construct, they collapse cleanly onto a single curve (see Figure 2.7). This enables us to analyze how parameters can compensate each other. For example, rather than viewing strong repression as a consequence of low IPTG concentration c or high repressor copy number R, we can now observe that strong repression is achieved when the free energy $F(c) \leq -5k_BT$, a condition which can be reached in a number of ways.

While our experiments validated the theoretical predictions in the case of simple

repression, we expect the framework presented here to apply much more generally to different biological instances of allosteric regulation. For example, we can use this model to study more complex systems such as when transcription factors interact with multiple operators [24]. We can further explore different regulatory configurations such as corepression, activation, and coactivation, each of which are found in *E. coli* (see Supplemental Section 2.15). This work can also serve as a springboard to characterize not just the mean but the full gene expression distribution and thus quantify the impact of noise on the system [59]. Another extension of this approach would be to theoretically predict and experimentally verify whether the repressor-inducer dissociation constants K_A and K_I or the energy difference $\Delta\varepsilon_{AI}$ between the allosteric states can be tuned by making single amino acid substitutions in the transcription factor [23, 26]. Finally, we expect that the kind of rigorous quantitative description of the allosteric phenomenon provided here will make it possible to construct biophysical models of fitness for allosteric proteins similar to those already invoked to explore the fitness effects of transcription factor binding site strengths and protein stability [60–62].

To conclude, we find that our application of the MWC model provides an accurate, predictive framework for understanding simple repression by allosteric transcription factors. To reach this conclusion, we analyzed the model in the context of a well-characterized system, in which each parameter had a clear biophysical meaning. As many of these parameters had been measured or inferred in previous studies, this gave us a minimal model with only two free parameters which we inferred from a single data set. We then accurately predicted the behavior of seventeen other data sets in which repressor copy number and repressor-DNA binding energy were systematically varied. In addition, our model allowed us to understand how key properties such as the leakiness, saturation, dynamic range, $[EC_{50}]$, and effective Hill coefficient depended upon the small set of parameters governing this system. Finally, we show that by framing inducible simple repression in terms of free energy, the data from all of our experimental strains collapse cleanly onto a single curve, illustrating the many ways in which a particular output can be targeted. In total, these results show that a thermodynamic formulation of the MWC model supersedes phenomenological fitting functions for understanding transcriptional regulation by allosteric proteins.

2.4 Methods

Bacterial Strains and DNA Constructs

All strains used in these experiments were derived from *E. coli* K12 MG1655 with the *lac* operon removed, adapted from those created and described in Ref. [9, 13]. Briefly, the operator variants and YFP reporter gene were cloned into a pZS25 background which contains a *lacUV5* promoter that drives expression as is shown schematically in Figure 2.2. These constructs carried a kanamycin resistance gene and were integrated into the *galK* locus of the chromosome using λ Red recombineering [63]. The *lacI* gene was constitutively expressed via a $P_{LtetO-1}$ promoter [53], with ribosomal binding site mutations made to vary the LacI copy number as described in Ref. [64] using site-directed mutagenesis (Quickchange II; Stratagene), with further details in Ref. [9]. These *lacI* constructs carried a chloramphenicol resistance gene and were integrated into the *ybcN* locus of the chromosome. Final strain construction was achieved by performing repeated P1 transduction [65] of the different operator and *lacI* constructs to generate each combination used in this work. Integration was confirmed by PCR amplification of the replaced chromosomal region and by sequencing. Primers and final strain genotypes are listed in Supplemental Section 2.16.

It is important to note that the rest of the *lac* operon (*lacZYA*) was never expressed. The LacY protein is a transmembrane protein which actively transports lactose as well as IPTG into the cell. As LacY was never produced in our strains, we assume that the extracellular and intracellular IPTG concentration was approximately equal due to diffusion across the membrane into the cell as is suggested by previous work [66].

To make this theory applicable to transcription factors with any number of DNA binding domains, we used a different definition for repressor copy number than has been used previously. We define the LacI copy number as the average number of repressor dimers per cell whereas in Ref. [9], the copy number is defined as the average number of repressor tetramers in each cell. To motivate this decision, we consider the fact that the LacI repressor molecule exists as a tetramer in *E. coli* [67] in which a single DNA binding domain is formed from dimerization of LacI proteins, so that wild-type LacI might be described as dimer of dimers. Since each dimer is allosterically independent (i.e. either dimer can be allosterically active or inactive, independent of the configuration of the other dimer) [31], a single LacI tetramer can be treated as two functional repressors. Therefore, we have simply

multiplied the number of repressors reported in Ref. [9] by a factor of two. This factor is included as a keyword argument in the numerous Python functions used to perform this analysis, as discussed in the code documentation.

A subset of strains in these experiments were measured using fluorescence microscopy for validation of the flow cytometry data and results. To aid in the high-fidelity segmentation of individual cells, the strains were modified to constitutively express an mCherry fluorophore. This reporter was cloned into a pZS4*1 backbone [53] in which mCherry is driven by the *lacUV5* promoter. All microscopy and flow cytometry experiments were performed using these strains.

Growth Conditions for Flow Cytometry Measurements

All measurements were performed with *E. coli* cells grown to mid-exponential phase in standard M9 minimal media (M9 5X Salts, Sigma-Aldrich M6030; 2 mM magnesium sulfate, Mallinckrodt Chemicals 6066-04; 100 μM calcium chloride, Fisher Chemicals C79-500) supplemented with 0.5% (w/v) glucose. Briefly, 500 μL cultures of *E. coli* were inoculated into Lysogeny Broth (LB Miller Powder, BD Medical) from a 50% glycerol frozen stock (-80°C) and were grown overnight in a 2 mL 96-deep-well plate sealed with a breathable nylon cover (Lab Pak—Nitex Nylon, Sefar America Inc. Cat. No. 241205) with rapid agitation for proper aeration. After approximately 12 to 15 hours, the cultures had reached saturation and were diluted 1000-fold into a second 2 mL 96-deep-well plate where each well contained 500 μL of M9 minimal media supplemented with 0.5% w/v glucose (anhydrous D-Glucose, Macron Chemicals) and the appropriate concentration of IPTG (Isopropyl β-D-1 thiogalactopyranoside Dioxane Free, Research Products International). These were sealed with a breathable cover and were allowed to grow for approximately eight hours. Cells were then diluted ten-fold into a round-bottom 96-well plate (Corning Cat. No. 3365) containing 90 μL of M9 minimal media supplemented with 0.5% w/v glucose along with the corresponding IPTG concentrations. For each IPTG concentration, a stock of 100-fold concentrated IPTG in double distilled water was prepared and partitioned into 100 μL aliquots. The same parent stock was used for all experiments described in this work.

Flow Cytometry

Unless explicitly mentioned, all fold-change measurements were collected on a Miltenyi Biotec MACSquant Analyzer 10 Flow Cytometer graciously provided by the Pamela Björkman lab at Caltech. Detailed information regarding the voltage

settings of the photo-multiplier detectors can be found in Table 2.1. Prior to each day's experiments, the analyzer was calibrated using MACSQuant Calibration Beads (Cat. No. 130-093-607) such that day-to-day experiments would be comparable. All YFP fluorescence measurements were collected via 488 nm laser excitation coupled with a 525/50 nm emission filter. Unless otherwise specified, all measurements were taken over the course of two to three hours using automated sampling from a 96-well plate kept at approximately 4° - 10°C on a MACS Chill 96 Rack (Cat. No. 130-094-459). Cells were diluted to a final concentration of approximately 4×10^4 cells per μL which corresponded to a flow rate of 2,000-6,000 measurements per second, and acquisition for each well was halted after 100,000 events were detected. Once completed, the data were extracted and immediately processed using the following methods.

Unsupervised Gating of Flow Cytometry Data

Flow cytometry data will frequently include a number of spurious events or other undesirable data points such as cell doublets and debris. The process of restricting the collected data set to those data determined to be "real" is commonly referred to as gating. These gates are typically drawn manually [68] and restrict the data set to those points which display a high degree of linear correlation between their forward-scatter (FSC) and side-scatter (SSC). The development of unbiased and unsupervised methods of drawing these gates is an active area of research [69, 70]. For our purposes, we assume that the fluorescence level of the population should be log-normally distributed about some mean value. With this assumption in place, we developed a method that allows us to restrict the data used to compute the mean fluorescence intensity of the population to the smallest two-dimensional region of the log(FSC) vs. log(SSC) space in which 40% of the data is found. This was performed by fitting a bivariate Gaussian distribution and restricting the data used for calculation to those that reside within the 40th percentile. This procedure is described in more detail in the supplementary information as well as in a Jupyter notebook located in this paper's Github repository (`https://rpgroup-pboc.github.io/mwc_induction/code/notebooks/unsupervised_gating.html`).

Experimental Determination of Fold-Change

For each strain and IPTG concentration, the fold-change in gene expression was calculated by taking the ratio of the population mean YFP expression in the presence of LacI repressor to that of the population mean in the absence of LacI repressor.

However, the measured fluorescence intensity of each cell also includes the autofluorescence contributed by the weak excitation of the myriad protein and small molecules within the cell. To correct for this background, we computed the fold change as

$$\text{fold-change} = \frac{\langle I_{R>0}\rangle - \langle I_{\text{auto}}\rangle}{\langle I_{R=0}\rangle - \langle I_{\text{auto}}\rangle}, \tag{2.13}$$

where $\langle I_{R>0}\rangle$ is the average cell YFP intensity in the presence of repressor, $\langle I_{R=0}\rangle$ is the average cell YFP intensity in the absence of repressor, and $\langle I_{\text{auto}}\rangle$ is the average cell autofluorescence intensity, as measured from cells that lack the *lac*-YFP construct.

Bayesian Parameter Estimation

In this work, we determine the the most likely parameter values for the inducer dissociation constants K_A and K_I of the active and inactive state, respectively, using Bayesian methods. We compute the probability distribution of the value of each parameter given the data D, which by Bayes' theorem is given by

$$P(K_A, K_I \mid D) = \frac{P(D \mid K_A, K_I)P(K_A, K_I)}{P(D)}, \tag{2.14}$$

where D is all the data composed of independent variables (repressor copy number R, repressor-DNA binding energy $\Delta\varepsilon_{RA}$, and inducer concentration c) and one dependent variable (experimental fold-change). $P(D \mid K_A, K_I)$ is the likelihood of having observed the data given the parameter values for the dissociation constants, $P(K_A, K_I)$ contains all the prior information on these parameters, and $P(D)$ serves as a normalization constant, which we can ignore in our parameter estimation. Equation 2.5 assumes a deterministic relationship between the parameters and the data, so in order to construct a probabilistic relationship as required by Equation 2.14, we assume that the experimental fold-change for the i^{th} datum given the parameters is of the form

$$\text{fold-change}^{(i)}_{\text{exp}} = \left(1 + \frac{\left(1 + \frac{c^{(i)}}{K_A}\right)^2}{\left(1 + \frac{c^{(i)}}{K_A}\right)^2 + e^{-\beta\Delta\varepsilon_{AI}}\left(1 + \frac{c^{(i)}}{K_I}\right)^2}\frac{R^{(i)}}{N_{NS}}e^{-\beta\Delta\varepsilon^{(i)}_{RA}}\right)^{-1} + \epsilon^{(i)}, \tag{2.15}$$

where $\epsilon^{(i)}$ represents the departure from the deterministic theoretical prediction for the i^{th} data point. If we assume that these $\epsilon^{(i)}$ errors are normally distributed with mean zero and standard deviation σ, the likelihood of the data given the parameters

is of the form

$$P(D|K_A, K_I, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \prod_{i=1}^{n} \exp\left[-\frac{(\text{fold-change}_{\text{exp}}^{(i)} - \text{fold-change}(K_A, K_I, R^{(i)}, \Delta\varepsilon_{RA}^{(i)}, c^{(i)}))^2}{2\sigma^2}\right], \tag{2.16}$$

where $\text{fold-change}_{\text{exp}}^{(i)}$ is the experimental fold-change and $\text{fold-change}(\cdots)$ is the theoretical prediction. The product $\prod_{i=1}^{n}$ captures the assumption that the n data points are independent. Note that the likelihood and prior terms now include the extra unknown parameter σ. In applying Equation 2.16, a choice of K_A and K_I that provides better agreement between theoretical fold-change predictions and experimental measurements will result in a more probable likelihood.

Both mathematically and numerically, it is convenient to define $\tilde{k}_A = -\log \frac{K_A}{1\,\text{M}}$ and $\tilde{k}_I = -\log \frac{K_I}{1\,\text{M}}$ and fit for these parameters on a log scale. Dissociation constants are scale invariant, so that a change from 10 μM to 1 μM leads to an equivalent increase in affinity as a change from 1 μM to 0.1 μM. With these definitions we assume for the prior $P(\tilde{k}_A, \tilde{k}_I, \sigma)$ that all three parameters are independent. In addition, we assume a uniform distribution for $\tilde{k}_A$ and $\tilde{k}_I$ and a Jeffreys prior [35] for the scale parameter σ. This yields the complete prior

$$P(\tilde{k}_A, \tilde{k}_I, \sigma) \equiv \frac{1}{(\tilde{k}_A^{\text{max}} - \tilde{k}_A^{\text{min}})} \frac{1}{(\tilde{k}_I^{\text{max}} - \tilde{k}_I^{\text{min}})} \frac{1}{\sigma}. \tag{2.17}$$

These priors are maximally uninformative meaning that they imply no prior knowledge of the parameter values. We defined the $\tilde{k}_A$ and $\tilde{k}_A$ ranges uniform on the range of -7 to 7, although we note that this particular choice does not affect the outcome provided the chosen range is sufficiently wide.

Putting all these terms together we can now sample from $P(\tilde{k}_A, \tilde{k}_I, \sigma \mid D)$ using Markov Chain Monte Carlo (see Github repository, `https://rpgroup-pboc.github.io/mwc_induction/code/notebooks/bayesian_parameter_estimation`) to compute the most likely parameter as well as the error bars (given by the 95% credible region) for K_A and K_I.

Data Curation

All of the data used in this work as well as all relevant code can be found at the dedicated website `http://rpgroup-pboc.github.io/mwc_induction`. Data were collected, stored, and preserved using the Git version control software in combination with off-site storage and hosting website GitHub. Code used to generate all figures and complete all processing step as and analyses are available on

the GitHub repository. Many analysis files are stored as instructive Jupyter Notebooks. The scientific community is invited to fork our repositories and open constructive issues on the Github repository `https://www.github.com/rpgroup-pboc/mwc_induction`.

Acknowledgements

This work has been a wonderful exercise in scientific collaboration. We thank Hernan Garcia for information and advice for working with these bacterial strains, Pamela Björkman and Rachel Galimidi for access and training for use of the Miltenyi Biotec MACSQuant flow cytometer, and Colin deBakker of Milteny Biotec for useful advice and instruction in flow cytometry. The experimental front of this work began at the Physiology summer course at the Marine Biological Laboratory in Woods Hole, MA operated by the University of Chicago. We thank Simon Alamos, Nalin Ratnayeke, and Shane McInally for their work on the project during the course. We also thank Suzannah Beeler, Justin Bois, Robert Brewster, Ido Golding, Soichi Hirokawa, Jané Kondev, Tom Kuhlman, Heun Jin Lee, Muir Morrison, Nigel Orme, Alvaro Sanchez, and Julie Theriot for useful advice and discussion. This work was supported by La Fondation Pierre-Gilles de Gennes, the Rosen Center at Caltech, and the National Institutes of Health DP1 OD000217 (Director's Pioneer Award), R01 GM085286, and 1R35 GM118043-01 (MIRA). Nathan Belliveau is a Howard Hughes Medical Institute International Student Research fellow.

2.5 Supplemental Information: Inferring Allosteric Parameters from Previous Data

The fold-change profile described by Equation 2.5 features three unknown parameters K_A, K_I, and $\Delta\varepsilon_{AI}$. In this section, we explore different conceptual approaches to determining these parameters. We first discuss how the induction titration profile of the simple repression constructs used in this paper are not sufficient to determine all three MWC parameters simultaneously, since multiple degenerate sets of parameters can produce the same fold-change response. We then utilize an additional data set from Ref. [10] to determine the parameter $\Delta\varepsilon_{AI} = 4.5\ k_BT$, after which the remaining parameters K_A and K_I can be extracted from any induction profile with no further degeneracy.

Degenerate Parameter Values

In this section, we discuss how multiple sets of parameters may yield identical fold-change profiles. More precisely, we shall show that if we try to fit the data in Figure 2.4C to the fold-change Equation 2.5 and extract the three unknown parameters (K_A, K_I, and $\Delta\varepsilon_{AI}$), then multiple degenerate parameter sets would yield equally good fits. In other words, this data set alone is insufficient to uniquely determine the actual physical parameter values of the system. This problem persists even when fitting multiple data sets simultaneously as in Supplemental Section 2.11.

In Figure 2.8A, we fit the $R = 260$ data by fixing $\Delta\varepsilon_{AI}$ to the value shown on the x-axis and determine the parameters K_A and K_I given this constraint. We use the fold-change function Equation 2.5 but with $\beta\Delta\varepsilon_{RA}$ modified to the form $\beta\Delta\tilde{\varepsilon}_{RA}$ in Equation 2.21 to account for the underlying assumptions used when fitting previous data (see Supplemental Section 2.5 for a full explanation of why this modification is needed).

The best-fit curves for several different values of $\Delta\varepsilon_{AI}$ are shown in Figure 2.8B. Note that these fold-change curves are nearly overlapping, demonstrating that different sets of parameters can yield nearly equivalent responses. Without more data, the relationships between the parameter values shown in Figure 2.8A represent the maximum information about the parameter values that can be extracted from the data. Additional experiments which independently measure any of these unknown parameters could resolve this degeneracy. For example, NMR measurements could be used to directly measure the fraction $(1 + e^{-\beta\Delta\varepsilon_{AI}})^{-1}$ of active repressors in the absence of IPTG [71, 72].

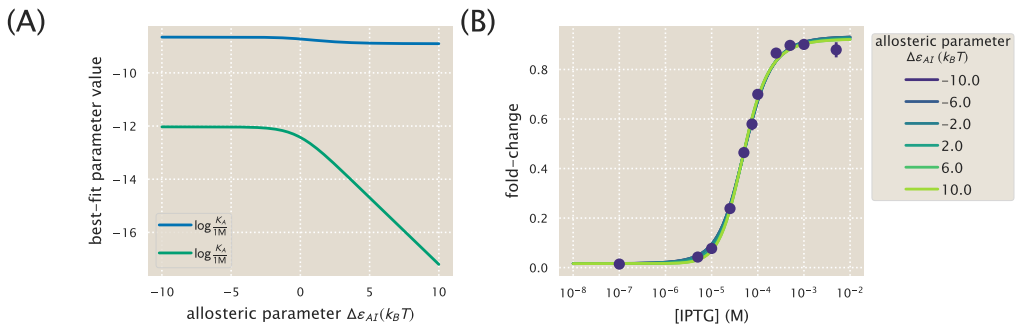

Figure 2.8: **Multiple sets of parameters yield identical fold-change responses.** (A) The data for the O2 strain ($\Delta\varepsilon_{RA} = -13.9\ k_BT$) with $R = 260$ in Figure 2.4C was fit using Equation 2.5 with $n = 2$. $\Delta\varepsilon_{AI}$ is forced to take on the value shown on the x-axis, while the K_A and K_I parameters are fit freely. (B) The resulting best-fit functions for several value of $\Delta\varepsilon_{AI}$ all yield nearly identical fold-change responses.

Computing $\Delta\varepsilon_{AI}$

As shown in the previous section, the fold-change response of a single strain is not sufficient to determine the three MWC parameters (K_A, K_I, and $\Delta\varepsilon_{AI}$), since degenerate sets of parameters yield nearly identical fold-change responses. To circumvent this degeneracy, we now turn to some previous data from the *lac* system in order to determine the value of $\Delta\varepsilon_{AI}$ in Equation 2.5 for the induction of the *lac*repressor. Specifically, we consider two previous sets of work from: (1) Ref. [9] and (2) Ref. [10], both of which measured fold-change with the same simple repression system in the absence of inducer ($c = 0$) but at various repressor copy numbers R. The original analysis for both data sets assumed that in the absence of inducer all of the *lac*repressors were in the active state. As a result, the effective binding energies they extracted were a convolution of the DNA binding energy $\Delta\varepsilon_{RA}$ and the allosteric energy difference $\Delta\varepsilon_{AI}$ between the *lac* repressor's active and inactive states. We refer to this convoluted energy value as $\Delta\tilde{\varepsilon}_{RA}$. We first disentangle the relationship between these parameters in Garcia and Phillips (Ref. [9]) and then use this relationship to extract the value of $\Delta\varepsilon_{AI}$ from the Brewster et al. dataset.

Garcia and Phillips determined the total repressor copy numbers R of different strains using quantitative Western blots. Then they measured the fold-change at these repressor copy numbers for simple repression constructs carrying the O1, O2, O3, and Oid *lac* operators integrated into the chromosome. These data were then fit to the following thermodynamic model to determine the repressor-DNA binding

energies $\Delta\tilde{\varepsilon}_{RA}$ for each operator,

$$\text{fold-change}(c = 0) = \left(1 + \frac{R}{N_{NS}} e^{-\beta\Delta\tilde{\varepsilon}_{RA}}\right)^{-1}. \tag{2.18}$$

Note that this functional form does not exactly match our fold-change Equation 2.5 in the limit $c = 0$,

$$\text{fold-change}(c = 0) = \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{2.19}$$

since it is missing the factor $\frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}}$ which specifies what fraction of repressors are in the active state in the absence of inducer,

$$\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} = p_A(0). \tag{2.20}$$

In other words, Garcia and Phillips assumed that in the absence of inducer, all repressors were active. In terms of our notation, the convoluted energy values $\Delta\tilde{\varepsilon}_{RA}$ extracted by Garcia and Phillips (namely, $\Delta\tilde{\varepsilon}_{RA} = -15.3\ k_BT$ for O1 and $\Delta\tilde{\varepsilon}_{RA} = -17.0\ k_BT$ for Oid) represent

$$\beta\Delta\tilde{\varepsilon}_{RA} = \beta\Delta\varepsilon_{RA} - \log\left(\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}}\right). \tag{2.21}$$

Note that if $e^{-\beta\Delta\varepsilon_{AI}} \ll 1$, then nearly all of the repressors are active in the absence of inducer so that $\Delta\tilde{\varepsilon}_{RA} \approx \Delta\varepsilon_{RA}$. In simple repression systems where we definitively know the value of $\Delta\varepsilon_{RA}$ and R, we can use Equation 2.19 to determine the value of $\Delta\varepsilon_{AI}$ by comparing with experimentally determined fold-change values. However, the binding energy values that we use from Ref. [9] are effective parameters $\Delta\tilde{\varepsilon}_{RA}$. In this case, we are faced with an undetermined system in which we have more variables than equations, and we are thus unable to determine the value of $\Delta\varepsilon_{AI}$. In order to obtain this parameter, we must turn to a more complex regulatory scenario which provides additional constraints that allow us to fit for $\Delta\varepsilon_{AI}$.

A variation on simple repression in which multiple copies of the promoter are available for repressor binding (for instance, when the simple repression construct is on plasmid) can be used to circumvent the problems that arise when using $\Delta\tilde{\varepsilon}_{RA}$. This is because the behavior of the system is distinctly different when the number of active repressors $p_A(0)R$ is less than or greater than the number of available promoters N. Repression data for plasmids with known copy number N allows us to perform a fit for the value of $\Delta\varepsilon_{AI}$.

To obtain an expression for a system with multiple promoters N, we follow Ref. [11], writing the fold-change in terms of the the grand canonical ensemble as

$$\text{fold-change} = \frac{1}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}}, \tag{2.22}$$

where $\lambda_r = e^{\beta\mu}$ is the fugacity and μ is the chemical potential of the repressor. The fugacity will enable us to easily enumerate the possible states available to the repressor.

To determine the value of λ_r, we first consider that the total number of repressors in the system, R_{tot}, is fixed and given by

$$R_{\text{tot}} = R_S + R_{NS}, \tag{2.23}$$

where R_S represents the number of repressors specifically bound to the promoter and R_{NS} represents the number of repressors nonspecifically bound throughout the genome. The value of R_S is given by

$$R_S = N \frac{\lambda_r e^{-\beta \Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}}, \tag{2.24}$$

where N is the number of available promoters in the cell. Note that in counting N, we do not distinguish between promoters that are on plasmid or chromosomally integrated provided that they both have the same repressor-operator binding energy [11]. The value of R_{NS} is similarly give by

$$R_{NS} = N_{NS} \frac{\lambda_r}{1 + \lambda_r}, \tag{2.25}$$

where N_{NS} is the number of non-specific sites in the cell (recall that we use $N_{NS} = 4.6 \times 10^6$ for *E. coli*).

Substituting in Equations 2.24 and 2.25 into the modified Equation 2.23 yields the form

$$p_A(0) R_{\text{tot}} = \frac{1}{1 + e^{-\beta \Delta\varepsilon_{AI}}} \left(N \frac{\lambda_r e^{-\beta \Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} \right), \tag{2.26}$$

where we recall from Equation 2.21 that $\beta \Delta\varepsilon_{RA} = \beta \Delta\tilde{\varepsilon}_{RA} + \log\left(\frac{1}{1+e^{-\beta \Delta\varepsilon_{AI}}}\right)$. Numerically solving for λ_r and plugging the value back into Equation 2.22 yields a fold-change function in which the only unknown parameter is $\Delta\varepsilon_{AI}$.

With these calculations in hand, we can now determine the value of the $\Delta\varepsilon_{AI}$ parameter. Figure 2.9A shows how different values of $\Delta\varepsilon_{AI}$ lead to significantly different

fold-change response curves. Thus, analyzing the specific fold-change response of any strain with a known plasmid copy number N will fix $\Delta\varepsilon_{AI}$. Interestingly, the inflection point of Equation 2.26 occurs near $p_A(0)R_{\text{tot}} = N$ (as shown by the triangles in Figure 2.9A), so that merely knowing where the fold-change response transitions from concave down to concave up is sufficient to obtain a rough value for $\Delta\varepsilon_{AI}$. We note, however, that for $\Delta\varepsilon_{AI} \gtrsim 5\ k_BT$, increasing $\Delta\varepsilon_{AI}$ further does not affect the fold-change because essentially every repressor will be in the active state in this regime. Thus, if the $\Delta\varepsilon_{AI}$ is in this regime, we can only bound it from below.

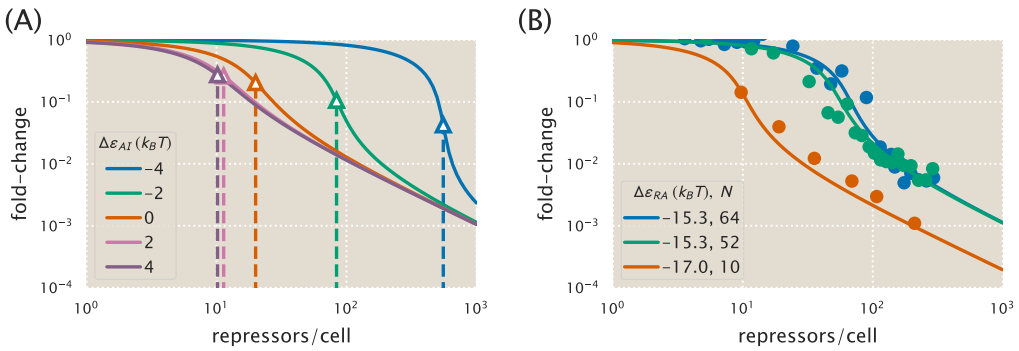

Figure 2.9: **Fold-change of multiple identical genes.** (A) In the presence of $N = 10$ identical promoters, the fold-change Equation 2.22 depends strongly on the allosteric energy difference $\Delta\varepsilon_{AI}$ between the *lac* repressor's active and inactive states. The vertical dotted lines represent the number of repressors at which $R_A = N$ for each value of $\Delta\varepsilon_{AI}$. (B) Using fold-change measurements from [10] for the operators and gene copy numbers shown, we can determine the most likely value $\Delta\varepsilon_{AI} = 4.5\ k_BT$ for LacI.

We now analyze experimental induction data for different strains with known plasmid copy numbers to determine $\Delta\varepsilon_{AI}$. Figure 2.9B shows experimental measurements of fold-change for two O1 promoters with $N = 64$ and $N = 52$ copy numbers and one Oid promoter with $N = 10$ from Ref. [10]. By fitting these data to Equation 2.22, we extracted the parameter value $\Delta\varepsilon_{AI} = 4.5\ k_BT$. Substituting this value into Equation 2.20 shows that 99% of the repressors are in the active state in the absence of inducer and $\Delta\tilde{\varepsilon}_{RA} \approx \Delta\varepsilon_{RA}$, so that all of the previous energies and calculations made by Refs. [9, 10] were accurate.

2.6 Supplemental Information: Induction of Simple Repression with Multiple Promoters or Competitor Sites

We made the choice to perform all of our experiments using strains in which a single copy of our simple repression construct had been integrated into the chromosome. This stands in contrast to the methods used by a number of other studies [4, 6, 26, 31, 33, 36, 39, 73], in which reporter constructs are placed on plasmid, meaning that the number of constructs in the cell is not precisely known. It is also common to express repressor on plasmid to boost its copy number, which results in an uncertain value for repressor copy number. Here we show that our treatment of the MWC model has broad predictive power beyond the single-promoter scenario we explore experimentally, and indeed can account for systems in which multiple promoters compete for the repressor of interest. Additionally, we demonstrate the importance of having precise control over these parameters, as they can have a significant effect on the induction profile.

Chemical Potential Formulation to Calculate Fold-Change

In this section, we discuss a simple repression construct which we generalize in two ways from the scenario discussed in the text. First, we will allow the repressor to bind to N_S identical specific promoters whose fold-change we are interested in measuring, with each promoter containing a single repressor binding site ($N_S = 1$ in the main text). Second, we consider N_C identical competitor sites which do not regulate the promoter of interest, but whose binding energies are substantially stronger than non-specific binding ($N_C = 0$ in the main text). As in the main text, we assume that the rest of the genome contains N_{NS} non-specific binding sites for the repressor. As in Supplemental Section 2.5, we can write the fold-change Equation 2.2 in the grand canonical ensemble as

$$\text{fold-change} = \frac{1}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}}, \tag{2.27}$$

where λ_r is the fugacity of the repressor and $\Delta\varepsilon_{RA}$ represents the energy difference between the repressor's binding affinity to the specific operator of interest relative to the repressor's non-specific binding affinity to the rest of the genome.

We now expand our definition of the total number of repressors in the system, R_{tot}, so that it is given by

$$R_{\text{tot}} = R_S + R_{NS} + R_C, \tag{2.28}$$

where R_S, R_{NS}, and R_C represent the number of repressors bound to the specific promoter, a non-specific binding site, or to a competitor binding site, respectively.

The value of R_S is given by

$$R_S = N_S \frac{\lambda_r e^{-\beta \Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}}, \tag{2.29}$$

where N_S is the number of specific binding sites in the cell. The value of R_{NS} is similarly give by

$$R_{NS} = N_{NS} \frac{\lambda_r}{1 + \lambda_r}, \tag{2.30}$$

where N_{NS} is the number of non-specific sites in the cell (recall that we use $N_{NS} = 4.6 \times 10^6$ for *E. coli*), and R_C is given by

$$R_C = N_C \frac{\lambda_r e^{-\beta \Delta\varepsilon_C}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_C}}, \tag{2.31}$$

where N_C is the number of competitor sites in the cell and $\Delta\varepsilon_C$ is the binding energy of the repressor to the competitor site relative to its non-specific binding energy to the rest of the genome.

To account for the induction of the repressor, we replace the total number of repressors R_{tot} in Equation 2.28 by the number of active repressors in the cell, $p_A(c)R_{\text{tot}}$. Here, p_A denotes the probability that the repressor is in the active state (Equation 2.4),

$$p_A(c) = \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta \Delta\varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n}. \tag{2.32}$$

Substituting in Equations 2.29, 2.30 and 2.31 into the modified Equation 2.28 yields the form

$$p_A(c)R_{\text{tot}} = N_S \frac{\lambda_r e^{-\beta \Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta \Delta\varepsilon_C}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_C}}. \tag{2.33}$$

For systems where the number of binding sites N_S, N_{NS}, and N_C are known, together with the binding affinities $\Delta\varepsilon_{RA}$ and $\Delta\varepsilon_C$, we can solve numerically for λ_r and then substitute it into Equation 2.27 to obtain a fold-change at any concentration of inducer c. In the following sections, we will theoretically explore the induction curves implied by Equation 2.33 for a number of different combinations of simple repression binding sites, thereby predicting how the system would behave if additional specific or competitor binding sites were introduced.

Variable Repressor Copy Number (R) with Multiple Specific Binding Sites ($N_S > 1$)

In the the main text, we consider the induction profiles of strains with varying R but a single, specific binding site $N_S = 1$ (see Figure 2.5). Here we predict the

induction profiles for similar strains in which R is varied, but $N_S > 1$, as shown in Figure 2.10. The top row shows induction profiles in which $N_S = 10$ and the bottom row shows profiles in which $N_S = 100$, assuming three different choices for the specific operator binding sites given by the O1, O2, and O3 operators. These values of N_S were chosen to mimic the common scenario in which a promoter construct is placed on either a low or high copy number plasmid. A few features stand out in these profiles. First, as the magnitude of N_S surpasses the number of repressors R, the leakiness begins to increase significantly, since there are no longer enough repressors to regulate all copies of the promoter of interest. Second, in the cases where $\Delta\varepsilon_{RA} = -15.3\ k_BT$ for the O1 operator or $\Delta\varepsilon_{RA} = -13.9\ k_BT$ for the O2 operator, the profiles where $N_S = 100$ are notably sharper than the profiles where $N_S = 10$, and it is possible to achieve dynamic ranges approaching 1. Finally, it is interesting to note that the profiles for the O3 operator where $\Delta\varepsilon_{RA} = -9.7\ k_BT$ are nearly indifferent to the value of N_S.

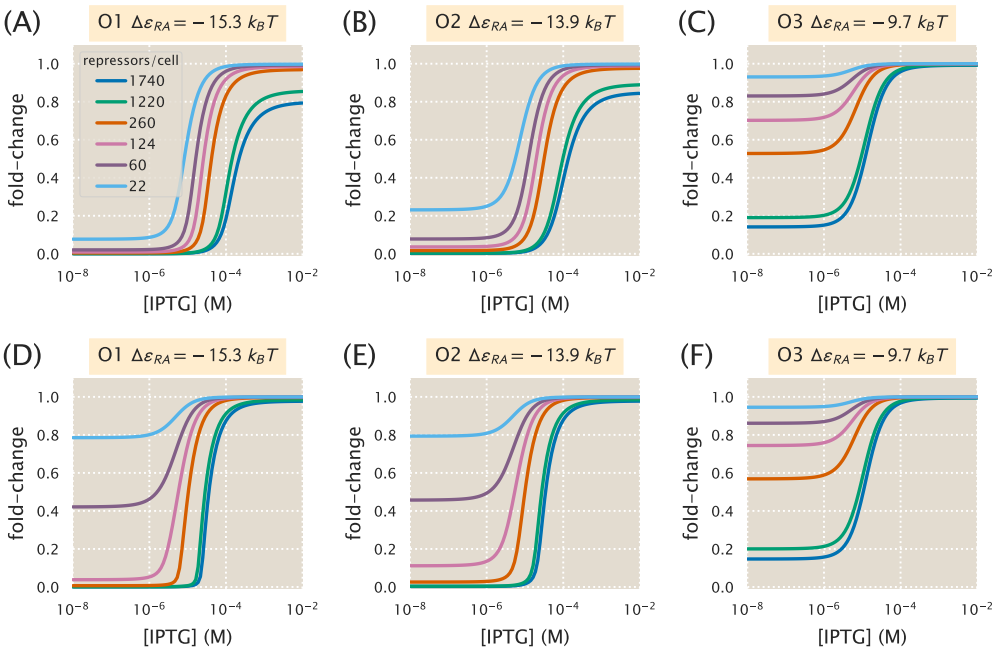

Figure 2.10: **Induction with variable *R* and multiple specific binding sites.** Induction profiles are shown for strains with variable R and $\Delta\varepsilon_{RA} = -15.3$, -13.9, or $-9.7\ k_BT$. (A-C) The number of specific sites, N_S, is held constant at 10 as R and $\Delta\varepsilon_{RA}$ are varied. (D-F) N_S is held constant at 100 as R and $\Delta\varepsilon_{RA}$ are varied. These situations mimic the common scenario in which a promoter construct is placed on either a low or high copy number plasmid.

Variable Number of Specific Binding Sites N_S with Fixed Repressor Copy Number (R)

The second set of scenarios we consider is the case in which the repressor copy number $R = 260$ is held constant while the number of specific promoters N_S is varied (see Figure 2.11). Again we see that leakiness is increased significantly when $N_S > R$, though all profiles for $\Delta\varepsilon_{RA} = -9.7\ k_BT$ exhibit high leakiness, making the effect less dramatic for this operator. Additionally, we find again that adjusting the number of specific sites can produce induction profiles with maximal dynamic ranges. In particular, the O1 and O2 profiles with $\Delta\varepsilon_{RA} = -15.3$ and $-13.9\ k_BT$, respectively, have dynamic ranges approaching 1 for $N_S = 50$ and 100.

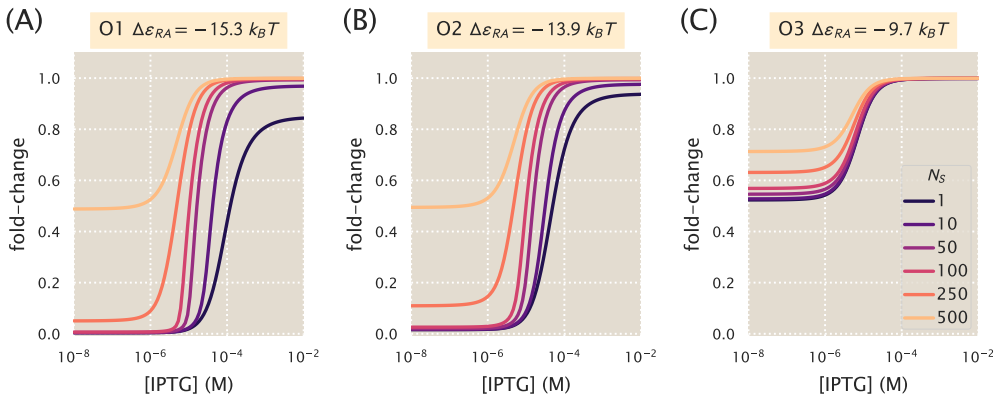

Figure 2.11: **Induction with variable specific sites and fixed R.** Induction profiles are shown for strains with $R = 260$ and (A) $\Delta\varepsilon_{RA} = -15.3\ k_BT$, (B) $\Delta\varepsilon_{RA} = -13.9\ k_BT$, or (C) $\Delta\varepsilon_{RA} = -9.7\ k_BT$. The number of specific sites N_S is varied from 1 to 500.

Competitor Binding Sites

An intriguing scenario is presented by the possibility of competitor sites elsewhere in the genome. This serves as a model for situations in which a promoter of interest is regulated by a transcription factor that has multiple targets. This is highly relevant, as the majority of transcription factors in *E. coli* have at least two known binding sites, with approximately 50 transcription factors having more than ten known binding sites [74, 75]. If the number of competitor sites and their average binding energy is known, however, they can be accounted for in the model. Here, we predict the induction profiles for strains in which $R = 260$ and $N_S = 1$, but there is a variable number of competitor sites N_C with a strong binding energy $\Delta\varepsilon_C = -17.0\ k_BT$. In the presence of such a strong competitor, when $N_C > R$ the leakiness is greatly increased, as many repressors are siphoned into the pool of competitor sites. This

is most dramatic for the case where $\Delta\varepsilon_{RA} = -9.7\ k_BT$, in which it appears that no repression occurs at all when $N_C = 500$. Interestingly, when $N_C < R$ the effects of the competitor are not especially notable.

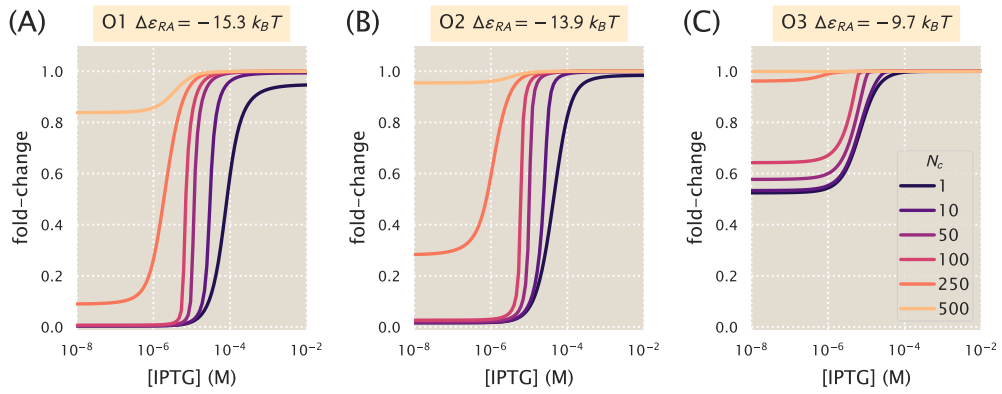

Figure 2.12: **Induction with variable competitor sites, a single specific site, and fixed *R*.** Induction profiles are shown for strains with $R = 260$, $N_s = 1$, and (A) $\Delta\varepsilon_{RA} = -15.3\ k_BT$ for the O1 operator, (B) $\Delta\varepsilon_{RA} = -13.9\ k_BT$ for the O2 operator, or (C) $\Delta\varepsilon_{RA} = -9.7\ k_BT$ for the O3 operator. The number of specific sites, N_C, is varied from 1 to 500. This mimics the common scenario in which a transcription factor has multiple binding sites in the genome.

Properties of the Induction Response

As discussed in the main body of the paper, our treatment of the MWC model allows us to predict key properties of induction responses. Here, we consider the leakiness, saturation, and dynamic range (see Figure 2.1) by numerically solving Equation 2.33 in the absence of inducer, $c = 0$, and in the presence of saturating inducer $c \to \infty$. Using Equation 2.32, the former case is given by

$$R_{\text{tot}} \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} = N_S \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta\Delta\varepsilon_C}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_C}}, \tag{2.34}$$

whereupon substituting in the value of λ_r into Equation 2.27 will yield the leakiness. Similarly, the limit of saturating inducer is found by determining λ_r from the form

$$R_{\text{tot}} \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^2} = N_S \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta\Delta\varepsilon_C}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_C}}. \tag{2.35}$$

In Figure 2.13 we show how the leakiness, saturation, and dynamic range vary with R and $\Delta\varepsilon_{RA}$ in systems with $N_S = 10$ or $N_S = 100$. An inflection point occurs where $N_S = R$, with leakiness and dynamic range behaving differently when

$R < N_S$ than when $R > N_S$. This transition is more dramatic for $N_S = 100$ than for $N_S = 10$. Interestingly, the saturation values consistently approach 1, indicating that full induction is easier to achieve when multiple specific sites are present. Moreover, dynamic range values for O1 and O2 strains with $\Delta\varepsilon_{RA} = -15.3$ and $-13.9\ k_BT$ approach 1 when $R > N_S$, although when $N_S = 10$ there is a slight downward dip owing to saturation values of less than 1 at high repressor copy numbers.

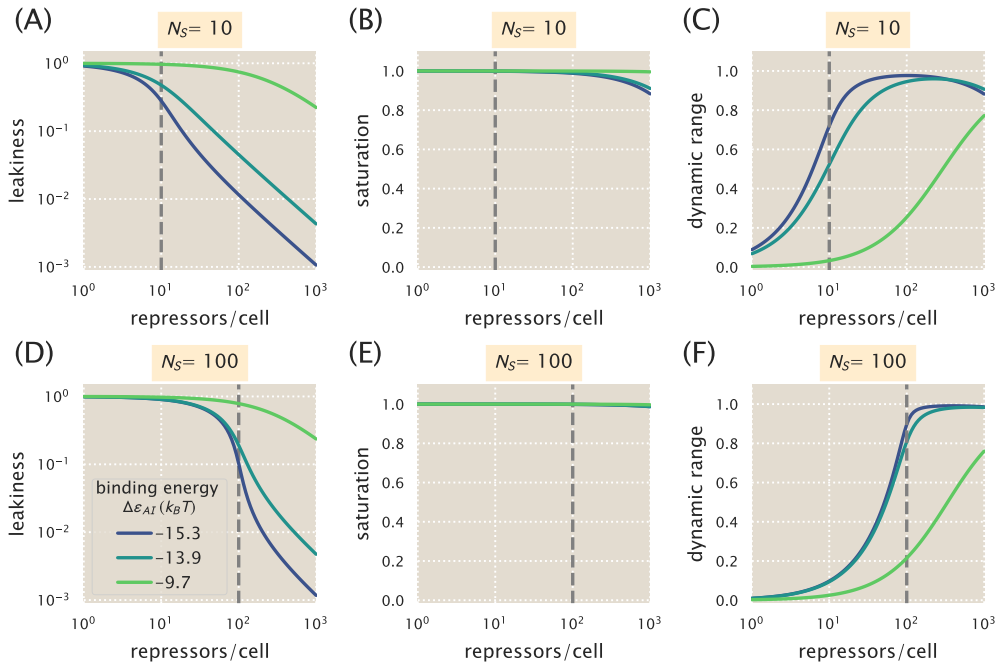

Figure 2.13: **Phenotypic properties of induction with multiple specific binding sites.** The leakiness (A, D), saturation (B, E), and dynamic range (C, F) are shown for systems with number of specific binding sites $N_S = 10$ (A-C) or $N_S = 100$ (D-F). The dashed vertical line indicates the point at which $N_S = R$.

In Figure 2.14 we similarly show how the leakiness, saturation, and dynamic range vary with R and $\Delta\varepsilon_{RA}$ in systems with $N_S = 1$ and multiple competitor sites $N_C = 10$ or $N_C = 100$. Each of the competitor sites has a binding energy of $\Delta\varepsilon_C = -17.0\ k_BT$. The phenotypic profiles are very similar to those for multiple specific sites shown in Figure 2.13, with sharper transitions at $R = N_C$ due to the greater binding strength of the competitor site. This indicates that introducing competitors has much the same effect on the induction phenotypes as introducing additional specific sites, as in either case the influence of the repressors is dampened when there are insufficient repressors to interact with all of the specific binding sites.

This section gives a quantitative analysis of the nuances imposed on induction

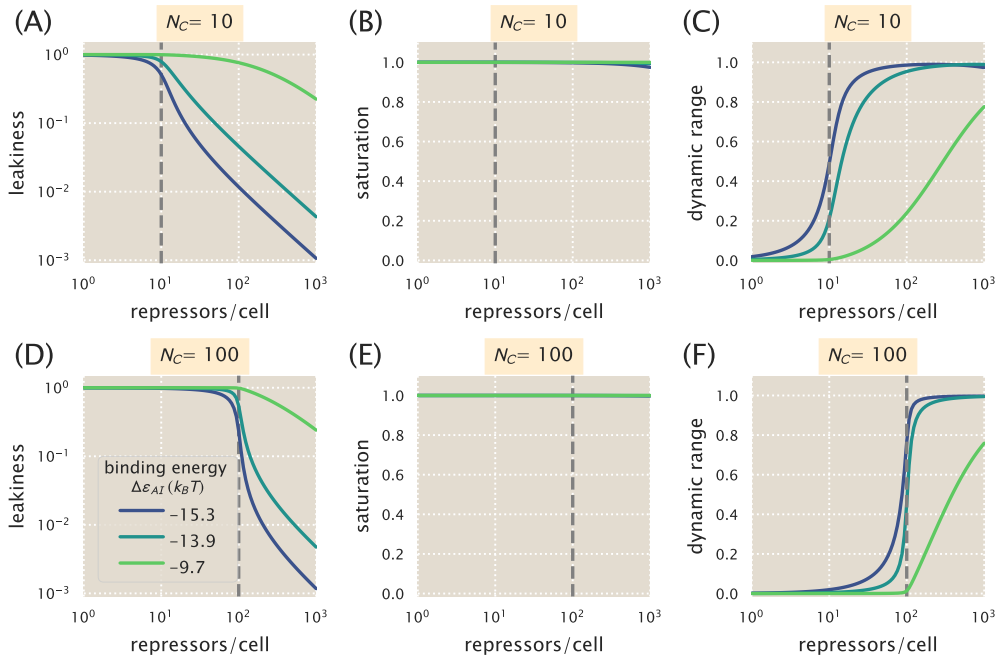

Figure 2.14: **Phenotypic properties of induction with a single specific site and multiple competitor sites.** The leakiness (A, D), saturation (B, E), and dynamic range (C, F) are shown for systems with a single specific binding site $N_S = 1$ and a number of competitor sites $N_C = 10$ (A-C) or $N_C = 100$ (D-F). All competitor sites have a binding energy of $\Delta\varepsilon_C = -17.0\ k_BT$. The dashed vertical line indicates the point at which $N_C = R$.

response in the case of systems involving multiple gene copies as are found in the vast majority of studies on induction. In these cases, the intrinsic parameters of the MWC model get entangled with the parameters describing gene copy number.

2.7 Supplemental Information: Flow Cytometry

In this section, we provide information regarding the equipment used to make experimental measurements of the fold-change in gene expression in the interests of transparency and reproducibility. We also provide a summary of our unsupervised method of gating the flow cytometry measurements for consistency between experimental runs.

Equipment

Due to past experience using the Miltenyi Biotec MACSQuant flow cytometer during the Physiology summer course at the Marine Biological Laboratory, we used the same flow cytometer for the formal measurements in this work graciously provided by the Pamela Björkman lab at Caltech. All measurements were made using an excitation wavelength of 488 nm with an emission filter set of 525/50 nm. This excitation wavelength provides approximately 40% of the maximum YFP absorbance [76], and this was found to be sufficient for the purposes of these experiments. A useful feature of modern flow cytometry is the high-sensitivity signal detection through the use of photomultiplier tubes (PMT) whose response can be tuned by adjusting the voltage. Thus, the voltage for the forward-scatter (FSC), side-scatter (SSC), and gene expression measurements were tuned manually to maximize the dynamic range between autofluorescence signal and maximal expression without losing the details of the population distribution. Once these voltages were determined, they were used for all subsequent measurements. Extremely low signal producing particles were discarded before data storage by setting a basal voltage threshold, thus removing the majority of spurious events. The various instrument settings for data collection are given in Table 2.1.

Table 2.1: **Instrument settings for data collection using the Miltenyi Biotec MACSQuant flow cytometer.** All experimental measurements were collected using these values.

Laser	Channel	Sensor Voltage
488 nm	Forward-Scatter (FSC)	423 V
488 nm	Side-Scatter (SSC)	537 V
488 nm	Intensity (B1 Filter, 525/50nm)	790 V
488 nm	Trigger (debris threshold)	24.5 V

Experimental Measurement

Prior to each day's experiments, the analyzer was calibrated using MACSQuant Calibration Beads (Cat. No. 130-093-607) such that day-to-day experiments would be comparable. A single data set consisted of seven bacterial strains, all sharing the same operator, with varying repressor copy numbers (R = 0, 22, 60, 124, 260, 1220, and 1740), in addition to an autofluorescent strain, under twelve IPTG concentrations. Data collection took place over two to three hours. During this time, the cultures were held at approximately 4°C by placing the 96-well plate on a MACSQuant ice block. Because the ice block thawed over the course of the experiment, the samples measured last were approximately at room temperature. This means that samples may have grown slightly by the end of the experiment. To confirm that this continued growth did not alter the measured results, a subset of experiments were run in reverse meaning that the fully induced cultures were measured first and the uninduced samples last. The plate arrangements and corresponding fold-change measurements are shown in Figure 2.15A and Figure 2.15B, respectively. The measured fold-change values in the reverse ordered plate appear to be drawn from the same distribution as those measured in the forward order, meaning that any growth that might have taken place during the experiment did not significantly affect the results. Both the forward and reverse data sets were used in our analysis.

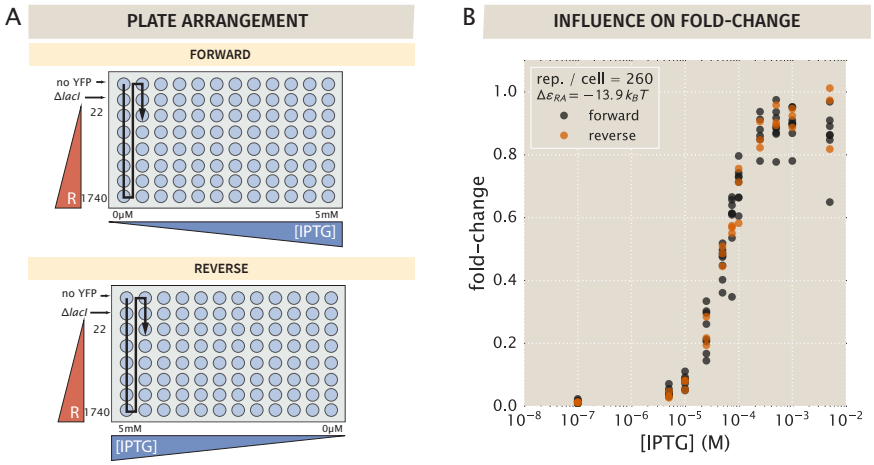

Figure 2.15: **Plate arrangements for flow cytometry.** (A) Samples were measured primarily in the forward arrangement with a subset of samples measured in reverse. The black arrow indicates the order in which samples were processed by the flow cytometer. (B) The experimentally measured fold-change values for the two sets of plate arrangements show that samples measured in the forward arrangement appear to be indistinguishable from those measured in reverse order.

Unsupervised Gating

Flow cytometry data will frequently include a number of spurious events or other undesirable data points such as cell doublets and debris. The process of restricting the collected data set to those data determined to be "real" is commonly referred to as gating. These gates are typically drawn manually [68] and restrict the data set to those points which display a high degree of linear correlation between their forward-scatter (FSC) and side-scatter (SSC). The development of unbiased and unsupervised methods of drawing these gates is an active area of research [69, 70].

For this study, we used an automatic unsupervised gating procedure to filter the flow cytometry data based on the front and side-scattering values returned by the MACSQuant flow cytometer. We assume that the region with highest density of points in these two channels corresponds to single-cell measurements. Everything extending outside of this region was discarded in order to exclude sources of error such as cell clustering, particulates, or other spurious events.

In order to define the gated region we fit a two-dimensional Gaussian function to the $\log_{10}$ forward-scattering (FSC) and the $\log_{10}$ side-scattering (SSC) data. We then kept a fraction $\alpha \in [0, 1]$ of the data by defining an elliptical region given by

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi^2_\alpha(p), \tag{2.36}$$

where $\boldsymbol{x}$ is the 2×1 vector containing the log(FSC) and log(SSC), $\boldsymbol{\mu}$ is the 2×1 vector representing the mean values of log(FSC) and log(SSC) as obtained from fitting a two-dimensional Gaussian to the data, and $\boldsymbol{\Sigma}$ is the 2×2 covariance matrix also obtained from the Gaussian fit. $\chi^2_\alpha(p)$ is the quantile function for probability p of the chi-squared distribution with two degrees of freedom. Figure 2.16 shows an example of different gating contours that would arise from different values of α in Equation 2.36. In this work, we chose $\alpha = 0.4$ which we deemed was a sufficient constraint to minimize the noise in the data. As explained in Supplemental Section 2.8 we compared our high throughput flow cytometry data with single cell microscopy, confirming that the automatic gating did not introduce systematic biases to the analysis pipeline. The specific code where this gating is implemented can be found in the Github repository, `https://github.com/RPGroup-PBoC/mwc_induction/blob/master/code/analysis/unsupervised_gating.ipynb`.

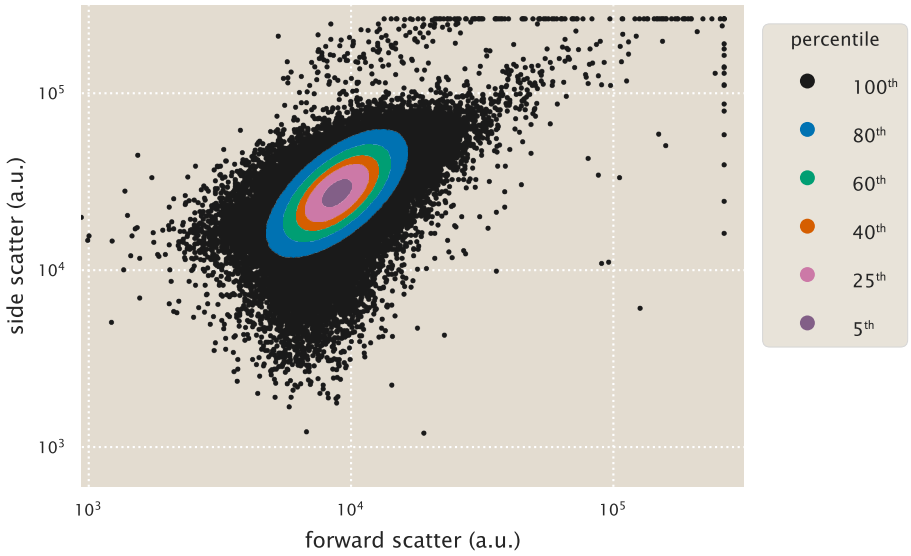

Figure 2.16: **Representative unsupervised gating contours.** Points indicate individual flow cytometry measurements of forward scatter and side scatter. Colored points indicate arbitrary gating contours ranging from 100% ($\alpha = 1.0$) to 5% ($\alpha = 0.05$). All measurements for this work were made computing the mean fluorescence from the 40^{th} percentile ($\alpha = 0.4$), shown as orange points.

Comparison of Flow Cytometry with Other Methods

Previous work from our lab experimentally determined fold-change for similar simple repression constructs using a variety of different measurement methods [10, 13]. Garcia and Phillips used the same background strains as the ones used in this work, but gene expression was measured with Miller assays based on colorimetric enzymatic reactions with the LacZ protein [9]. Ref. [10] used a LacI dimer with the tetramerization region replaced with an mCherry tag, where the fold-change was measured as the ratio of the gene expression rate rather than a single snapshot of the gene output.

Figure 2.17 shows the comparison of these methods along with the flow cytometry method used in this work. The consistency of these three readouts validates the quantitative use of flow cytometry and unsupervised gating to determine the fold-change in gene expression. However, one important caveat revealed by this figure is that the sensitivity of flow cytometer measurements is not sufficient to accurately determine the fold-change for the high repressor copy number strains in O1 without induction. Instead, a method with a large dynamic range such as the Miller assay is needed to accurately resolve the fold-change at such low expression levels.

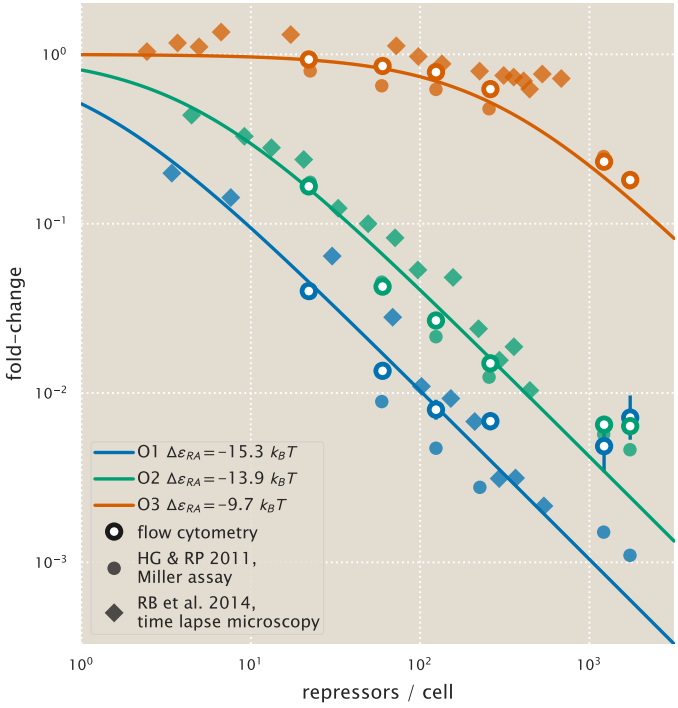

Figure 2.17: **Comparison of experimental methods to determine the fold-change.** The fold-change in gene expression for equivalent simple-repression constructs has been determined using three independent methods: flow cytometry (this work), colorimetric Miller assays [9], and video microscopy [10]. All three methods give consistent results, although flow cytometry measurements lose accuracy for fold-change less than 10^{-2}. Note that the repressor-DNA binding energies $\Delta\varepsilon_{RA}$ used for the theoretical predictions were determined in Ref. [9].

2.8 Supplemental Information: Single-Cell Microscopy

In this section, we detail the procedures and results from single-cell microscopy verification of our flow cytometry measurements. Our previous measurements of fold-change in gene expression have been measured using bulk-scale Miller assays [9] or through single-cell microscopy [10]. In this work, flow cytometry was an attractive method due to the ability to screen through many different strains at different concentrations of inducer in a short amount of time. To verify our results from flow cytometry, we examined two bacterial strains with different repressor-DNA binding energies ($\Delta\varepsilon_{RA}$) of $-13.9\ k_BT$ and $-15.3\ k_BT$ with $R = 260$ repressors per cell using fluorescence microscopy and estimated the values of the parameters K_A and K_I for direct comparison between the two methods. For a detailed explanation of the Python code implementation of the processing steps described below, please see this paper's Github repository, `https://rpgroup-pboc.github.io/mwc_induction/code/notebooks/unsupervised_gating.html`. An outline of our microscopy workflow can be seen in Figure 2.18.

Strains and Growth Conditions

Cells were grown in an identical manner to those used for measurement via flow cytometry (see Methods). Briefly, cells were grown overnight (between 10 and 13 hours) to saturation in rich media broth (LB) with $100\ \mu\text{g} \cdot \text{mL}^{-1}$ spectinomycin in a deep-well 96 well plate at 37°C. These cultures were then diluted 1000-fold into 500 μL of M9 minimal medium supplemented with 0.5% glucose and the appropriate concentration of the inducer IPTG. Strains were allowed to grow at 37°C with vigorous aeration for approximately 8 hours. Prior to mounting for microscopy, the cultures were diluted 10-fold into M9 glucose minimal medium in the absence of IPTG. Each construct was measured using the same range of inducer concentration values as was performed in the flow cytometry measurements (between 100 nM and 5 mM IPTG). Each condition was measured in triplicate in microscopy whereas approximately ten measurements were made using flow cytometry.

Imaging Procedure

During the last hour of cell growth, an agarose mounting substrate was prepared containing the appropriate concentration of the IPTG inducer. This mounting substrate was composed of M9 minimal medium supplemented with 0.5% glucose and 2% agarose (Life Technologies UltraPure Agarose, Cat. No. 16500100). This solution was heated in a microwave until molten followed by addition of the IPTG

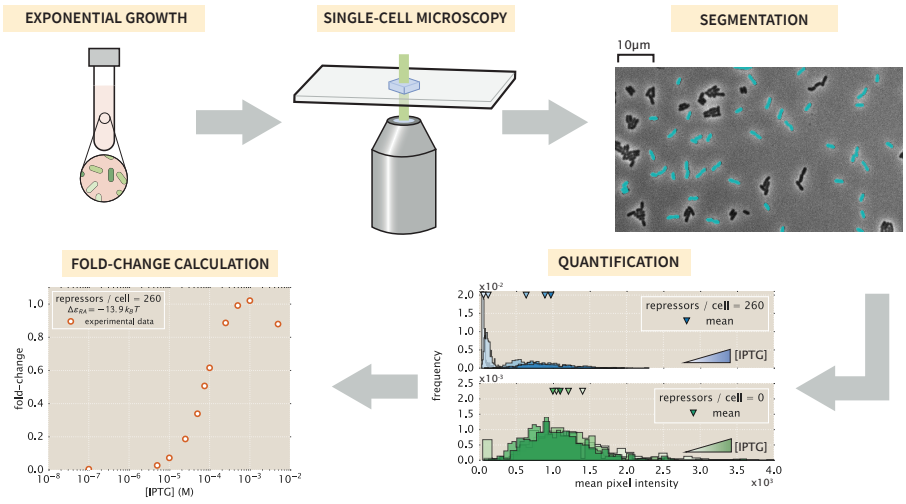

Figure 2.18: **Experimental workflow for single-cell microscopy**. For comparison with the flow cytometry results, the cells were grown in an identical manner to those described in the main text. Once cells had reached mid to late exponential growth, the cultures were diluted and placed on agarose substrates and imaged under 100× magnification. Regions of interest representing cellular mass were segmented and average single-cell intensities were computed. The means of the distributions were used to compute the fold-change in gene expression.

to the appropriate final concentration. This solution was then thoroughly mixed and a 500 μL aliquot was sandwiched between two glass coverslips and was allowed to solidify.

Once solid, the agarose substrates were cut into approximately 10 mm × 10 mm squares. An aliquot of one to two microliters of the diluted cell suspension was then added to each pad. For each concentration of inducer, a sample of the autofluorescence control, the $\Delta lacI$ constitutive expression control, and the experimental strain was prepared yielding a total of thirty-six agarose mounts per experiment. These samples were then mounted onto two glass-bottom dishes (Ted Pella Wilco Dish, Cat. No. 14027-20) and sealed with parafilm.

All imaging was performed on a Nikon Ti-Eclipse inverted fluorescent microscope outfitted with a custom-built laser illumination system and operated by the open-source MicroManager control software [77]. The YFP fluorescence was imaged using a CrystaLaser 514 nm excitation laser coupled with a laser-optimized (Semrock Cat. No. LF514-C-000) emission filter.

For each sample, between fifteen and twenty positions were imaged allowing for measurement of several hundred cells. At each position, a phase contrast image,

an mCherry image, and a YFP image were collected in that order with exposures on a time scale of ten to twenty milliseconds. For each channel, the same exposure time was used across all samples in a given experiment. All images were collected and stored in `ome.tiff` format. All microscopy images are available on the CaltechDATA online repository under DOI: 10.22002/D1.229.

Image Processing

Correcting Uneven Illumination

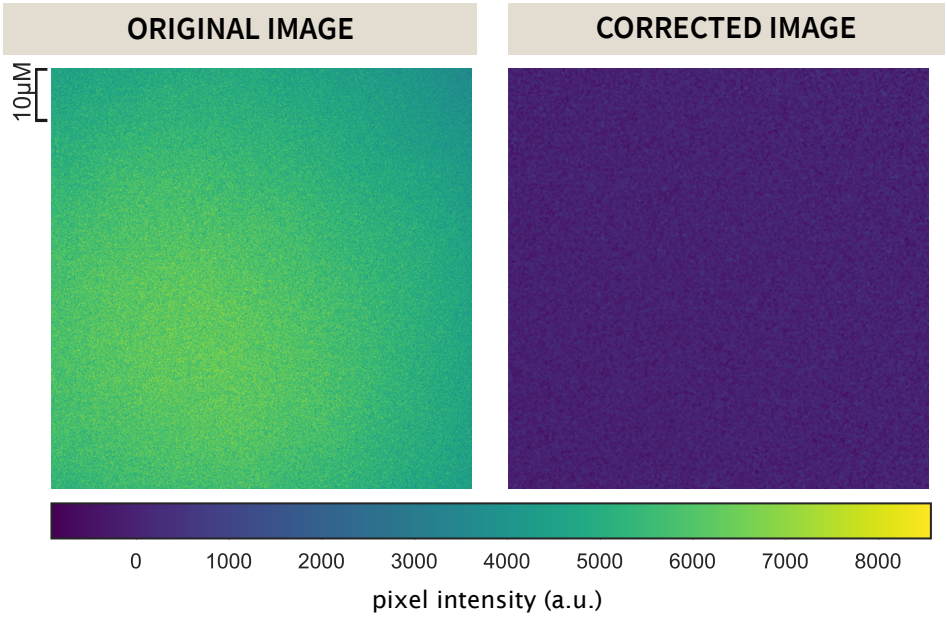

Figure 2.19: **Correction for uneven illumination.** A representative image of the illumination profile of the 512 nm excitation beam on a homogeneously fluorescent slide is shown in the left panel. This is corrected for using Equation 2.37 and is shown in the right panel.

The excitation laser has a two-dimensional gaussian profile. To minimize non-uniform illumination of a single field of view, the excitation beam was expanded to illuminate an area larger than that of the camera sensor. While this allowed for an entire field of view to be illuminated, there was still approximately a 10% difference in illumination across both dimensions. This nonuniformity was corrected for in post-processing by capturing twenty images of a homogeneously fluorescent plastic slide (Autofluorescent Plastic Slides, Chroma Cat. No. 920001) and averaging to generate a map of illumination intensity at any pixel I_{YFP}. To correct for shot noise in the camera (Andor iXon+ 897 EMCCD), twenty images were captured in the absence of illumination using the exposure time used for the experimental data.

Averaging over these images produced a map of background noise at any pixel I_{dark}. To perform the correction, each fluorescent image in the experimental acquisition was renormalized with respect to these average maps as

$$I_{\text{flat}} = \frac{I - I_{\text{dark}}}{I_{\text{YFP}} - I_{\text{dark}}} \langle I_{\text{YFP}} - I_{\text{dark}} \rangle, \tag{2.37}$$

where I_{flat} is the renormalized image and I is the original fluorescence image. An example of this correction can be seen in Figure 2.19.

Cell Segmentation

Each bacterial strain constitutively expressed an mCherry fluorophore from a low copy-number plasmid. This served as a volume marker of cell mass allowing us to segment individual cells through edge detection in fluorescence. We used the Marr-Hildreth edge detector [78] which identifies edges by taking the second derivative of a lightly Gaussian blurred image. Edges are identified as those regions which cross from highly negative to highly positive values or vice-versa within a specified neighborhood. Bacterial cells were defined as regions within an intact and closed identified edge. All segmented objects were then labeled and passed through a series of filtering steps.

To ensure that primarily single cells were segmented, we imposed area and eccentricity bounds. We assumed that single cells projected into two dimensions are roughly 2 μm long and 1 μm wide, so that cells are likely to have an area between 0.5 μm^2 and 6 μm. To determine the eccentricity bounds, we assumed that the a single cell can be approximated by an ellipse with semi-major (a) and semi-minor (b) axis lengths of 0.5 μm and 0.25 μm, respectively. The eccentricity of this hypothetical cell can be computed as

$$\text{eccentricity} = \sqrt{1 - \left(\frac{b}{a}\right)^2}, \tag{2.38}$$

yielding a value of approximately 0.8. Any objects with an eccentricity below this value were not considered to be single cells. After imposing both an area (Figure 2.20A) and eccentricity filter (Figure 2.20B), the remaining objects were considered cells of interest (Figure 2.20C) and the mean fluorescence intensity of each cell was extracted.

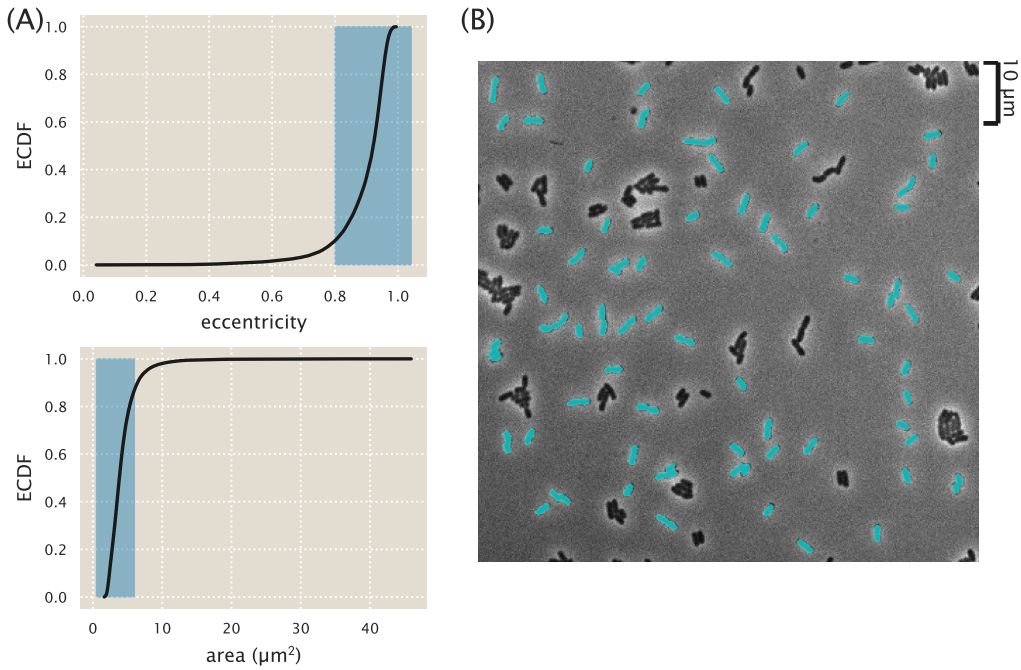

Figure 2.20: **Segmentation of single bacterial cells.** (A) Objects were selected if they had an eccentricity greater than 0.8 and an area between 0.5 μm^2 and 6 μm^2. Highlighted in blue are the regions considered to be representative of single cells. The black lines correspond to the empirical cumulative distribution functions for the parameter of interest. (B) A representative final segmentation mask is shown in which segmented cells are depicted in cyan over the phase contrast image.

Calculation of Fold-Change

Cells exhibited background fluorescence even in the absence of an expressed fluorophore. We corrected for this autofluorescence contribution to the fold-change calculation by subtracting the mean YFP fluorescence of cells expressing only the mCherry volume marker from each experimental measurement. The fold-change in gene expression was therefore calculated as

$$\text{fold-change} = \frac{\langle I_{R>0} \rangle - \langle I_{\text{auto}} \rangle}{\langle I_{R=0} \rangle - \langle I_{\text{auto}} \rangle}, \tag{2.39}$$

where $\langle I_{R>0} \rangle$ is the mean fluorescence intensity of cells expressing LacI, $\langle I_{\text{auto}} \rangle$ is the mean intensity of cells expressing only the mCherry volume marker, and $\langle I_{R=0} \rangle$ is the mean fluorescence intensity of cells in the absence of LacI. These fold-change values were very similar to those obtained through flow cytometry and were well described using the thermodynamic parameters used in the main text. With these experimentally measured fold-change values, the best-fit parameter values of the model were inferred and compared to those obtained from flow cytometry.

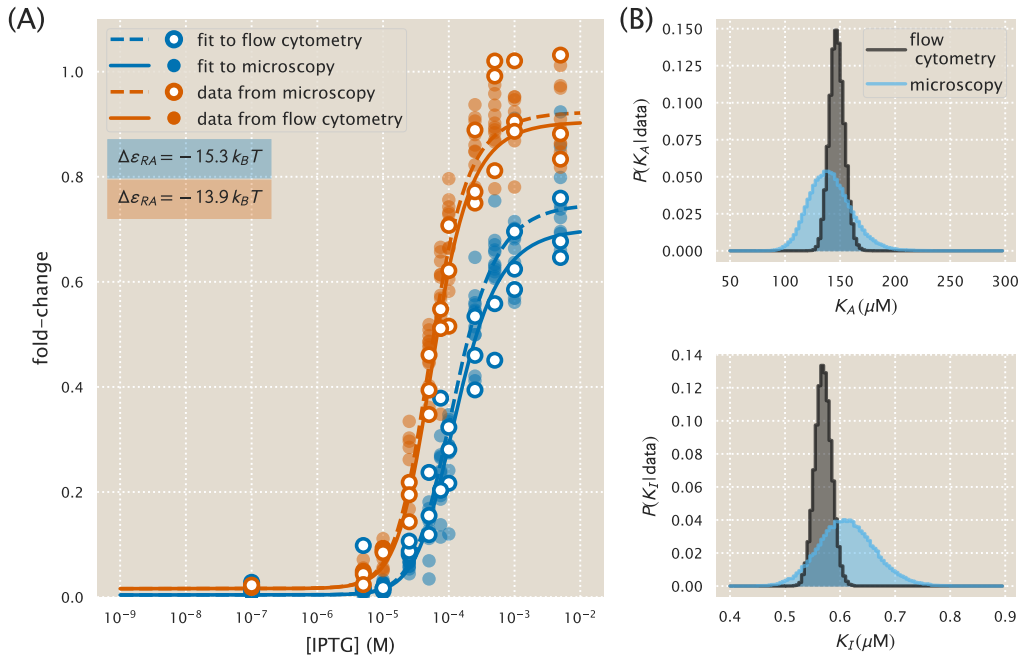

Figure 2.21: **Comparison of measured fold-change between flow cytometry and single-cell microscopy.** (A) Experimentally measured fold-change values obtained through single-cell microscopy and flow cytometry are shown as white filled and solid colored circles, respectively. Solid and dashed lines indicate the predicted behavior using the most likely parameter values of K_A and K_I inferred from flow cytometry data and microscopy data, respectively. The red and blue plotting elements correspond to the different operators O1 and O2 with binding energies $\Delta\varepsilon_{RA}$ of $-13.9\ k_BT$ and $-15.3\ k_BT$, respectively [9]. (B) The marginalized posterior distributions for K_A and K_I are shown in the top and bottom panel, respectively. The posterior distribution determined using the microscopy data is wider than that computed using the flow cytometry data due to a smaller fig collection of data sets (three for microscopy and ten for flow cytometry).

To confirm quantitative consistency between flow cytometry and microscopy, the parameter values of K_A and K_I were also estimated from three biological replicates of IPTG titration curves obtained by microscopy for strains with $R = 260$ and operators O1 and O2. Figure 2.21A shows the data from these measurements (orange circles) and the ten biological replicates from our flow cytometry measurements (blue circles), along with the fold-change predictions from each inference. In comparison with the values obtained by flow cytometry, each parameter estimate overlapped with the 95% credible region of our flow cytometry estimates, as shown in Figure 2.21B. Specifically, these values were $K_A = 142^{+40}_{-34}\ \mu$M and $K_I = 0.6^{+0.1}_{-0.1}\ \mu$M from microscopy and $K_A = 149^{+14}_{-12}\ \mu$M and $K_I = 0.57^{+0.03}_{-0.02}\ \mu$M from the flow

cytometry data. We note that the credible regions from the microscopy data shown in Figure 2.21B are much broader than those from flow cytometry due to the fewer number of replicates performed.

2.9 Supplemental Information: Fold-Change Sensitivity Analysis

In Figure 2.5 we found that the width of the credible regions varied widely depending on the repressor copy number R and repressor operator binding energy $\Delta\varepsilon_{RA}$. More precisely, the credible regions were much narrower for low repressor copy numbers R and weak binding energy $\Delta\varepsilon_{RA}$. In this section, we explain how this behavior comes about. We focus our attention on the maximum fold-change in the presence of saturating inducer given by Equation 2.7. While it is straightforward to consider the width of the credible regions at any other inducer concentration, Figure 2.5 shows that the credible regions are widest at saturation.

The width of the credible regions corresponds to how sensitive the fold-change is to the fit values of the dissociation constants K_A and K_I. To be quantitative, we define

$$\Delta\text{fold-change}_{K_A} \equiv \text{fold-change}(K_A, K_I^{\text{fit}}) - \text{fold-change}(K_A^{\text{fit}}, K_I^{\text{fit}}), \tag{2.40}$$

the difference between the fold-change at a particular K_A value relative to the best-fit dissociation constant $K_A^{\text{fit}} = 139 \times 10^{-6}\,\text{M}$. For simplicity, we keep the inactive state dissociation constant fixed at its best-fit value $K_I^{\text{fit}} = 0.53 \times 10^{-6}\,\text{M}$. A larger difference $\Delta\text{fold-change}_{K_A}$ implies a wider credible region. Similarly, we define the analogous quantity

$$\Delta\text{fold-change}_{K_I} = \text{fold-change}(K_A^{\text{fit}}, K_I) - \text{fold-change}(K_A^{\text{fit}}, K_I^{\text{fit}}) \tag{2.41}$$

to measure the sensitivity of the fold-change to K_I at a fixed K_A^{fit}. Figure 2.22 shows both of these quantities in the limit $c \to \infty$ for different repressor-DNA binding energies $\Delta\varepsilon_{RA}$ and repressor copy numbers R. See our Github repository (`https://github.com/RPGroup-PBoC/mwc_induction/blob/master/code/analysis/sensitivity_analysis.ipynb`) for the code that reproduces these plots.

To understand how the width of the credible region scales with $\Delta\varepsilon_{RA}$ and R, we can Taylor expand the difference in fold-change to first order, $\Delta\text{fold-change}_{K_A} \approx \frac{\partial\text{fold-change}}{\partial K_A}\left(K_A - K_A^{\text{fit}}\right)$, where the partial derivative has the form

$$\frac{\partial\text{fold-change}}{\partial K_A} = \frac{e^{-\beta\Delta\varepsilon_{AI}}\frac{n}{K_I}\left(\frac{K_A}{K_I}\right)^{n-1}}{\left(1 + e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n\right)^2}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-2}. \tag{2.42}$$

Similarly, the Taylor expansion $\Delta\text{fold-change}_{K_I} \approx \frac{\partial\text{fold-change}}{\partial K_I}\left(K_I - K_I^{\text{fit}}\right)$ features the partial derivative

$$\frac{\partial\text{fold-change}}{\partial K_I} = -\frac{e^{-\beta\Delta\varepsilon_{AI}}\frac{n}{K_I}\left(\frac{K_A}{K_I}\right)^n}{\left(1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n\right)^2}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\left(1+\frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-2}. \tag{2.43}$$

From Equations 2.42 and 2.43, we find that both $\Delta\text{fold-change}_{K_A}$ and $\Delta\text{fold-change}_{K_I}$ increase in magnitude with R and decrease in magnitude with $\Delta\varepsilon_{RA}$. Accordingly, we expect that the O3 strains (with the least negative $\Delta\varepsilon_{RA}$) and the strains with the smallest repressor copy number will lead to partial derivatives with smaller magnitude and hence to tighter credible regions. Indeed, this prediction is carried out in Figure 2.22.

Lastly, we note that Equations 2.42 and 2.43 enable us to quantify the scaling relationship between the width of the credible region and the two quantities R and $\Delta\varepsilon_{RA}$. For example, for the O3 strains, where the fold-change at saturating inducer concentration is ≈ 1, the right-most term in both equations which equals the fold-change squared is roughly 1. Therefore, we find that both $\frac{\partial\text{fold-change}}{\partial K_A}$ and $\frac{\partial\text{fold-change}}{\partial K_I}$ scale linearly with R and $e^{-\beta\Delta\varepsilon_{RA}}$. Thus the width of the $R = 22$ strain will be roughly 1/1000 as large as that of the $R = 1740$ strain; similarly, the width of the O3 curves will be roughly 1/1000 the width of the O1 curves.

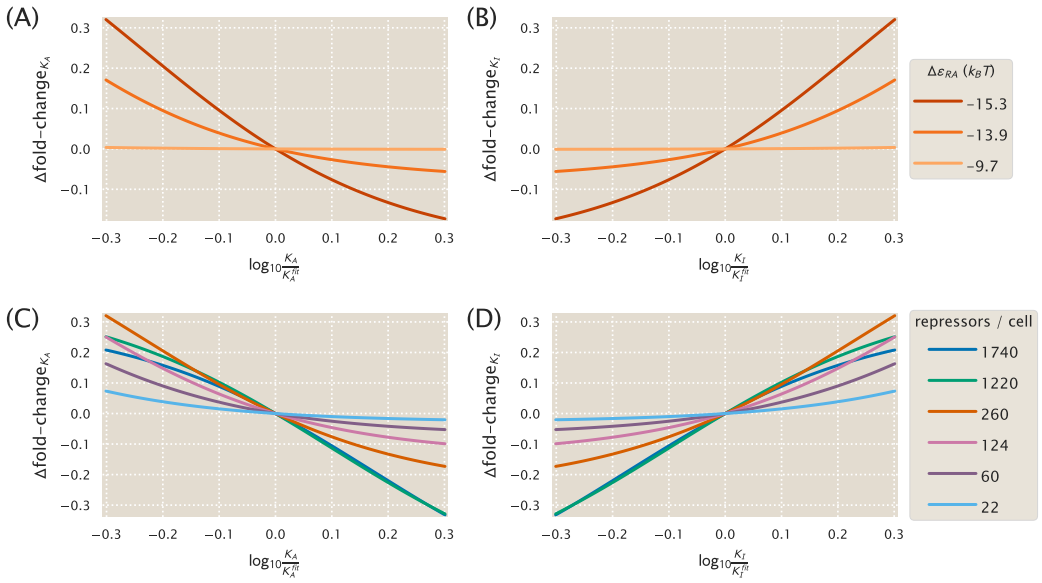

Figure 2.22: **Determining how sensitive the fold-change values are to the fit values of the dissociation constants.**(A) The difference $\Delta\text{fold-change}_{K_A}$ in fold change when the dissociation constant K_A is slightly offset from its best-fit value $K_A = 139^{+29}_{-22} \times 10^{-6}$ M, as given by Equation 2.40. Fold-change is computed in the limit of saturating inducer concentration ($c \rightarrow \infty$, see Equation 2.7) where the credible regions in Figure 2.5 are widest. The O3 strain ($\Delta\varepsilon_{RA} = -9.7\ k_BT$) is about 1/1000 as sensitive as the O1 operator to perturbations in the parameter values, and hence its credible region is roughly 1/1000 as wide. All curves were made using $R = 260$. (B) As in Panel A, but plotting the sensitivity of fold-change to the K_I parameter relative to the best-fit value $K_I = 0.53^{+0.04}_{-0.04} \times 10^{-6}$ M. Note that only the magnitude, and not the sign, of this difference describes the sensitivity of each parameter. Hence, the O3 strain is again less sensitive than the O1 and O2 strains. (C) As in Panel A, but showing how the fold-change sensitivity for different repressor copy numbers. The strains with lower repressor copy number are less sensitive to changes in the dissociation constants, and hence their corresponding curves in Figure 2.5 have tighter credible regions. All curves were made using $\Delta\varepsilon_{RA} = -13.9\ k_BT$. (D) As in Panel C, the sensitivity of fold-change with respect to K_I is again smallest (in magnitude) for the low repressor copy number strains.

2.10 Supplemental Information: Alternate Characterizations of Induction

In this section we discuss a different way to describe the induction data, namely, through using the conventional Hill approach. We first demonstrate how using a Hill function to characterize a single induction curve enables us to extract features (such as the midpoint and sharpness) of that single response, but precludes any predictions of the other seventeen strains. We then discuss how a thermodynamic model of simple repression coupled with a Hill approach to the induction response can both characterize an induction profile and predict the response of all eighteen strains, although we argue that such a description provides no insight into the allosteric nature of the protein and how mutations to the repressor would affect induction. We conclude the section by discussing the differences between such a model and the statistical mechanical model used in the main text.

Fitting Induction Curves using a Hill Function Approach

The Hill equation is a phenomenological function commonly used to describe data with a sigmoidal profile [7, 30, 32]. Its simplicity and ability to estimate the cooperativity of a system (through the Hill coefficient) has led to its widespread use in many domains of biology [79]. Nevertheless, the Hill function is often criticized as a physically unrealistic model and the extracted Hill coefficient is often difficult to contextualize in the physics of a system [80]. In the present work, we note that a Hill function, even if it is only used because of its simplicity, presents no mechanism to understand how a regulatory system's behavior will change if physical parameters such as repressor copy number or operator binding energy are varied. In addition, the Hill equation provides no foundation to explore how mutating the repressor (e.g., at its inducer-binding interface) would modify its induction profile, although statistical mechanical models have proved capable of characterizing such scenarios [42, 43, 45].

Consider the general Hill equation for a single induction profile given by

$$\text{fold-change} = (\text{leakiness}) + (\text{dynamic range})\frac{\left(\frac{c}{K}\right)^n}{1 + \left(\frac{c}{K}\right)^n}, \tag{2.44}$$

where, as in the main text, the leakiness represents the minimum fold-change, the dynamic range represents the difference between the maximum and minimum fold-change, K is the repressor-inducer dissociation constant, and n denotes the Hill coefficient that characterizes the sharpness of the curve ($n > 1$ signifies positive cooperativity, $n = 1$ denotes no cooperativity, and $n < 1$ represents negative coop-

erativity). Figure 2.23 shows how the individual induction profiles can be fit (using the same Bayesian methods as described in Supplemental Section 2.11) to this Hill response, yielding a similar response to that shown in Figure 2.4D. However, characterizing the induction response in this manner is unsatisfactory because each curve must be fit independently, thus removing our predictive power for other repressor copy numbers and binding sites.

The fitted parameters obtained from this approach are shown in Figure 2.24. These are rather unsatisfactory because they do not clearly reflect the properties of the physical system under consideration. For example, the dissociation constant K between LacI and inducer should not be affected by either the copy number of the repressor or the DNA binding energy, and yet we see upward trends as R is increased or the binding energy is decreased. Here, the K parameter ultimately describes the midpoint of the induction curve and therefore cannot strictly be considered a dissociation constant. Similarly, the Hill coefficient n does not directly represent the cooperativity between the repressor and the inducer as the molecular details of the copy number and DNA binding strength are subsumed in this parameter as well. While the leakiness and dynamic range describe important phenotypic properties of the induction response, this Hill approach leaves us with no means to predict them for other strains. In summary, the Hill equation Equation 2.44 cannot predict how an induction profile varies with repressor copy number, operator binding energy, or how mutations will alter the induction profile. To that end, we turn to a more sophisticated approach where we use the Hill function to describe the available fraction of repressor as a function of inducer concentration.

Fitting Induction Curves using a Combination Thermodynamic Model and Hill Function Approach

Motivated by the inability in the previous section to characterize all eighteen strains using the Hill function with a single set of parameters, here we combine the Hill approach with a thermodynamic model of simple repression to garner predictive power. More specifically, we will use the thermodynamic model in Figure 2.2A but substitute the statistical model in Figure 2.2B with the phenomenological Hill function Equation 2.44.

Following Equations 2.1, 2.2, and 2.3, fold-change is given by

$$\text{fold-change} = \left(1 + p_A(c)\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{2.45}$$

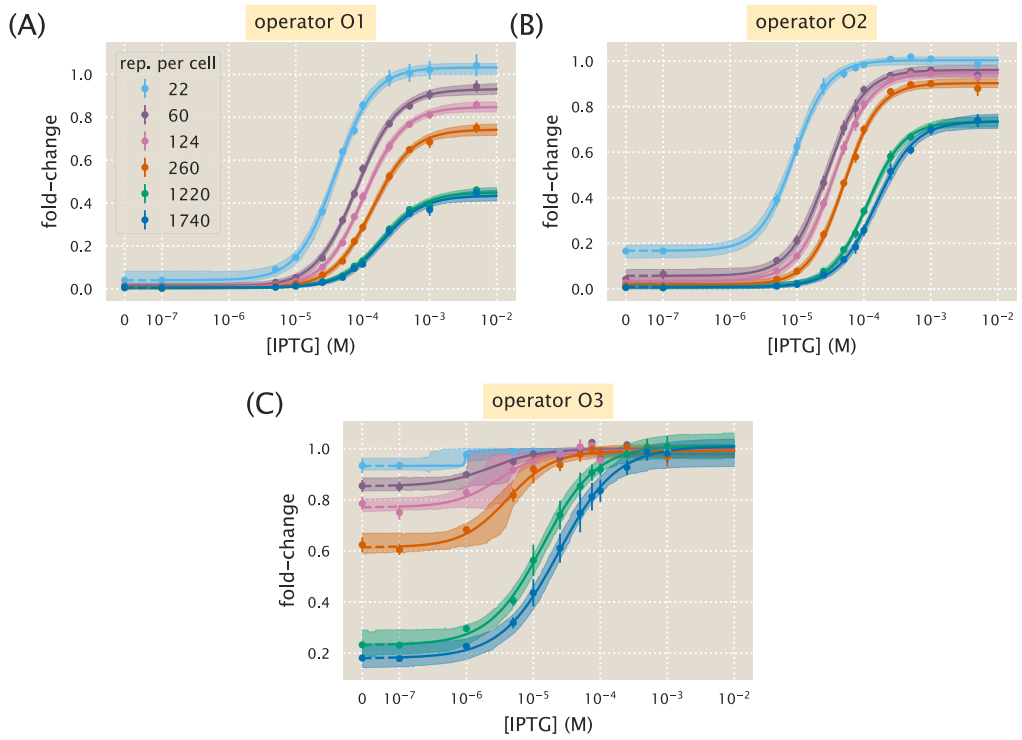

Figure 2.23: **Hill function and MWC analysis of each induction profile.** Data for each individual strain was fit to the general Hill function in Equation 2.44. (A) strains with O1 binding site, (B) strains with O2 binding site, and (C) strains with O3 binding site. Shaded regions indicate the bounds of the 95% credible region.

where the Hill function

$$p_A(c) = p_A^{\text{max}} - p_A^{\text{range}} \frac{\left(\frac{c}{K_D}\right)^n}{1 + \left(\frac{c}{K_D}\right)^n} \tag{2.46}$$

represents the fraction of repressors in the allosterically active state, with p_A^{max} denoting the fraction of active repressors in the absence of inducer and $p_A^{\text{max}} - p_A^{\text{range}}$ the minimum fraction of active repressors in the presence of saturating inducer. The Hill function characterizes the inducer-repressor binding while the thermodynamic model with the known constants R, N_{NS}, and $\Delta\varepsilon_{RA}$ describes how the induction profile changes with repressor copy number and repressor-operator binding energy.

As in the main text, we can fit the four Hill parameters—the vertical shift and stretch parameters p_A^{max} and p_A^{range}, the Hill coefficient n, and the inducer-repressor dissociation constant K_D—for a single induction curve and then use the fully characterized Equation 2.45 to describe the response of each of the eighteen strains. Figure 2.25 shows this process carried out by fitting the O2 $R = 260$ strain (white circles in Panel B) and predicting the behavior of the remaining seventeen strains.

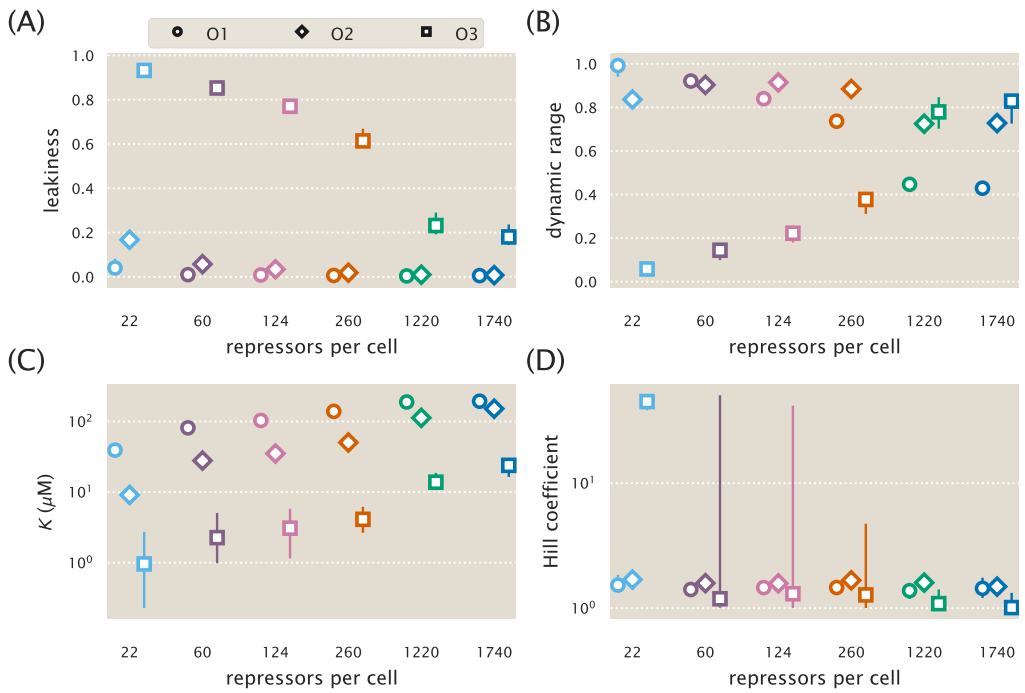

Figure 2.24: **Parameter values for the Hill equation fit to each individual titration.** The resulting fit parameters from the Hill function fits of Figure 2.23 are summarized. The large parameter intervals for many of the O3 strains are due to the flatter induction profile (as seen by its smaller dynamic range), and the ability for a large range of K and n values to describe the data.

Although the curves in Figure 2.25 are nearly identical to those in Figure 2.4 (which were made using the MWC model Equation 2.5), we stress that the Hill function approach is more complex than the MWC model (containing four parameters instead of three) and it obscures the relationships to the physical parameters of the system. For example, it is not clear whether the fit parameter $K_D = 4^{+2}_{-1} \times 10^{-6}$ M relays the dissociation constant between the inducer and active-state repressor, between the inducer and the inactive-state repressor, or some mix of the two quantities.

In addition, the MWC model Equation 2.5 naturally suggests further quantitative tests for the fold-change relationship. For example, mutating the repressor's inducer binding site would likely alter the repressor-inducer dissociation constants K_A and K_I, and it would be interesting to find out if such mutations also modify the allosteric energy difference $\Delta\varepsilon_{AI}$ between the repressor's active and inactive conformations. For our purposes, the Hill function Equation 2.46 falls short of the connection to the physics of the system and provides no intuition about how transcription depends upon such mutations. For these reasons, we present the thermodynamic model

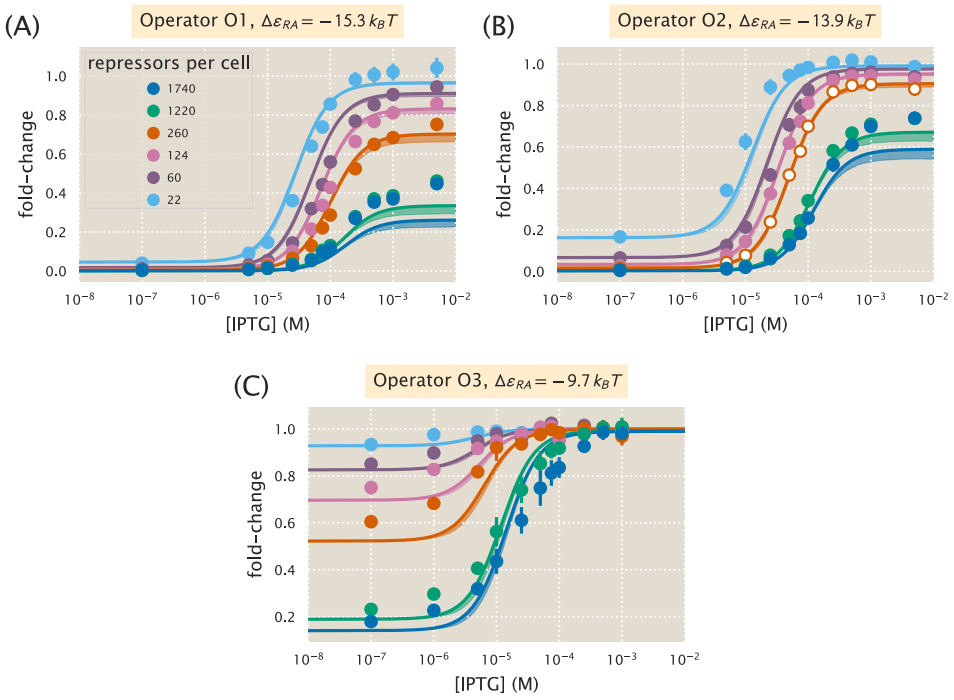

Figure 2.25: **A thermodynamic model coupled with a Hill analysis can characterize induction.** Combining a thermodynamic model of simple repression with the Hill function to characterize the repressor-inducer binding successfully characterizes the induction profiles of all eighteen strains. As in the main text, data was only fit for the O2 $R = 260$ strain using Equations 2.45 and 2.46 and the parameters $p_A^{\max} = 0.90^{+0.03}_{-0.01}$, $p_A^{\text{range}} = -0.90^{+0.02}_{-0.03}$, $n = 1.6^{+0.2}_{-0.1}$, and $K_D = 4^{+2}_{-1} \times 10^{-6}$ M. Shaded regions indicate bounds of the 95% credible region.

coupled with the statistical mechanical MWC model approach in the paper.

2.11 Supplemental Information: Global Fit of All Parameters

In the main text, we used the repressor copy numbers R and repressor-DNA binding energies $\Delta\varepsilon_{RA}$ as reported by Ref. [9]. However, any error in these previous measurements of R and $\Delta\varepsilon_{RA}$ will necessarily propagate into our own fold-change predictions. In this section we take an alternative approach to fitting the physical parameters of the system to that used in the main text. First, rather than fitting only a single strain, we fit the entire data set in Figure 2.5 along with microscopy data for the synthetic operator Oid (see Supplemental Section 2.12). In addition, we also simultaneously fit the parameters R and $\Delta\varepsilon_{RA}$ using the prior information given by the previous measurements. By using the entire data set and fitting all of the parameters, we obtain the best possible characterization of the statistical mechanical parameters of the system given our current state of knowledge. As a point of reference, we state all of the parameters of the MWC model derived in the text in Table 2.2.

To fit all of the parameters simultaneously, we follow a similar approach to the one detailed in the Methods section. Briefly, we perform a Bayesian parameter estimation of the dissociation constants K_A and K_I, the six different repressor copy numbers R corresponding to the six *lacI* ribosomal binding sites used in our work, and the four different binding energies $\Delta\varepsilon_{RA}$ characterizing the four distinct operators used to make the experimental strains. As in the main text, we fit the logarithms $\tilde{k}_A = -\log \frac{K_A}{1\,\mathrm{M}}$ and $\tilde{k}_I = -\log \frac{K_I}{1\,\mathrm{M}}$ of the dissociation constants, which grants better numerical stability.

As in Equation 2.15 and 2.16, we assume that deviations of the experimental fold-change from the theoretical predictions are normally distributed with mean zero and standard deviation σ. We begin by writing Bayes' theorem,

$$P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma \mid D) = \frac{P(D \mid \tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma) P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma)}{P(D)}, \tag{2.47}$$

where $\boldsymbol{R}$ is an array containing the six different repressor copy numbers to be fit, $\boldsymbol{\Delta\varepsilon_{RA}}$ is an array containing the four binding energies to be fit, and D is the experimental fold-change data. The term $P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma \mid D)$ gives the probability distributions of all of the parameters given the data. The term $P(D \mid \tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma)$ represents the likelihood of having observed our experimental data given some value for each parameter. $P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma)$ contains all the prior information on the values of these parameters. Lastly, $P(D)$ serves as a normalization constant and hence can be ignored.

Given n independent measurements of the fold-change, the first term in Equation 2.47 can be written as

$$P(D \mid \tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \prod_{i=1}^{n} \exp\left[-\frac{(\mathrm{fc}_{\mathrm{exp}}^{(i)} - \mathrm{fc}(\tilde{k}_A, \tilde{k}_I, R^{(i)}, \Delta\varepsilon_{RA}^{(i)}, c^{(i)}))^2}{2\sigma^2}\right], \tag{2.48}$$

where $\mathrm{fc}_{\mathrm{exp}}^{(i)}$ is the i^{th} experimental fold-change and $\mathrm{fc}(\cdots)$ is the theoretical prediction. Note that the standard deviation σ of this distribution is not known and hence needs to be included as a parameter to be fit.

The second term in Equation 2.47 represents the prior information of the parameter values. We assume that all parameters are independent of each other, so that

$$P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma) = P(\tilde{k}_A) \cdot P(\tilde{k}_I) \cdot \prod_i P(R^{(i)}) \cdot \prod_j P(\Delta\varepsilon_{RA}^{(j)}) \cdot P(\sigma), \tag{2.49}$$

where the superscript (i) indicates the repressor copy number of index i and the superscript (j) denotes the binding energy of index j. As above, we note that a prior must also be included for the unknown parameter σ.

Because we knew nothing about the values of $\tilde{k}_A$, $\tilde{k}_I$, and σ before performing the experiment, we assign maximally uninformative priors to each of these parameters. More specifically, we assign uniform priors to $\tilde{k}_A$ and $\tilde{k}_I$ and a Jeffreys prior to σ, indicating that K_A, K_I, and σ are scale parameters [35]. We do, however, have prior information for the repressor copy numbers and the repressor-DNA binding energies from Ref. [9]. This prior knowledge is included within our model using an informative prior for these two parameters, which we assume to be Gaussian. Hence each of the $R^{(i)}$ repressor copy numbers to be fit satisfies

$$P(R^{(i)}) = \frac{1}{\sqrt{2\pi\sigma_{R_i}^2}} \exp\left(-\frac{(R^{(i)} - \bar{R}^{(i)})^2}{2\sigma_{R_i}^2}\right), \tag{2.50}$$

where $\bar{R}^{(i)}$ is the mean repressor copy number and σ_{R_i} is the variability associated with this parameter as reported in Ref. [9]. Note that we use the given value of σ_{R_i} from previous measurements rather than leaving this as a free parameter.

Similarly, the binding energies $\Delta\varepsilon_{RA}^{(j)}$ are also assumed to have a Gaussian informative prior of the same form. We write it as

$$P(\Delta\varepsilon_{RA}^{(j)}) = \frac{1}{\sqrt{2\pi\sigma_{\varepsilon_j}^2}} \exp\left(-\frac{(\Delta\varepsilon_{RA}^{(j)} - \Delta\bar{\varepsilon}_{RA}^{(j)})^2}{2\sigma_{\varepsilon_j}^2}\right), \tag{2.51}$$

where $\Delta\bar{\varepsilon}_{RA}^{(j)}$ is the binding energy and σ_{ε_j} is the variability associated with that parameter around the mean value as reported in Ref. [9] .

The σ_{R_i} and σ_{ε_j} parameters will constrain the range of values for $R^{(i)}$ and $\Delta\varepsilon_{RA}^{(j)}$ found from the fitting. For example, if for some i the standard deviation σ_{R_i} is very small, it implies a strong confidence in the previously reported value. Mathematically, the exponential in Equation 2.50 will ensure that the best-fit $R^{(i)}$ lies within a few standard deviations of $\bar{R}^{(i)}$. Since we are interested in exploring which values could give the best fit, the errors are taken to be wide enough to allow the parameter estimation to freely explore parameter space in the vicinity of the best estimates. Putting all these terms together, we use Markov Chain Monte Carlo to sample the posterior distribution $P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma \mid D)$, enabling us to determine both the most likely value for each physical parameter as well as its associated credible region (see the Github repository for the implementation, `https://rpgroup-pboc.github.io/mwc_induction/code/notebooks/global_fits.html`).

Figure 2.26 shows the result of this global fit. When compared with Figure 2.5 we can see that fitting for the binding energies and the repressor copy numbers improves the agreement between the theory and the data. Table 2.3 summarizes the values of the parameters as obtained with this MCMC parameter inference. We note that even though we allowed the repressor copy numbers and repressor-DNA binding energies to vary, the resulting fit values were very close to the previously reported values. The fit values of the repressor copy numbers were all within one standard deviation of the previous reported values provided in Ref. [9]. And although some of the repressor-DNA binding energies differed by a few standard deviations from the reported values, the differences were always less than 1 k_BT, which represents a small change in the biological scales we are considering. The biggest discrepancy between our fit values and the previous measurements arose for the synthetic Oid operator, which we discuss in more detail in Supplemental Section 2.12.

Figure 2.27 shows the same key properties as in Figure 2.6, but uses the parameters obtained from this global fitting approach. We note that even by increasing the number of degrees of freedom in our fit, the result does not change substantially due to only minor improvements between the theoretical curves and data. For the O3 operator data, again, agreement between the predicted $[EC_{50}]$ and the effective Hill coefficient remains poor due the theory being unable to capture the steepness of the response curves.

Table 2.2: **Key model parameters for induction of an allosteric repressor.**

Parameter	Description
c	Concentration of the inducer
K_A, K_I	Dissociation constant between an inducer and the repressor in the active/inactive state
$\Delta\varepsilon_{AI}$	The difference between the free energy of repressor in the inactive and active states
$\Delta\varepsilon_P$	Binding energy between the RNAP and its specific binding site
$\Delta\varepsilon_{RA}, \Delta\varepsilon_{RI}$	Binding energy between the operator and the active/inactive repressor
n	Number of inducer binding sites per repressor
P	Number of RNAP
R_A, R_I, R	Number of active/inactive/total repressors
$p_A = \frac{R_A}{R}$	Probability that a repressor will be in the active state
p_{bound}	Probability that an RNAP is bound to the promoter of interest, assumed to be proportional to gene expression
fold-change	Ratio of gene expression in the presence of repressor to that in the absence of repressor
F	Free energy of the system
N_{NS}	The number of non-specific binding sites for the repressor in the genome
$\beta = \frac{1}{k_B T}$	The inverse product of the Boltzmann constant k_B and the temperature T of the system

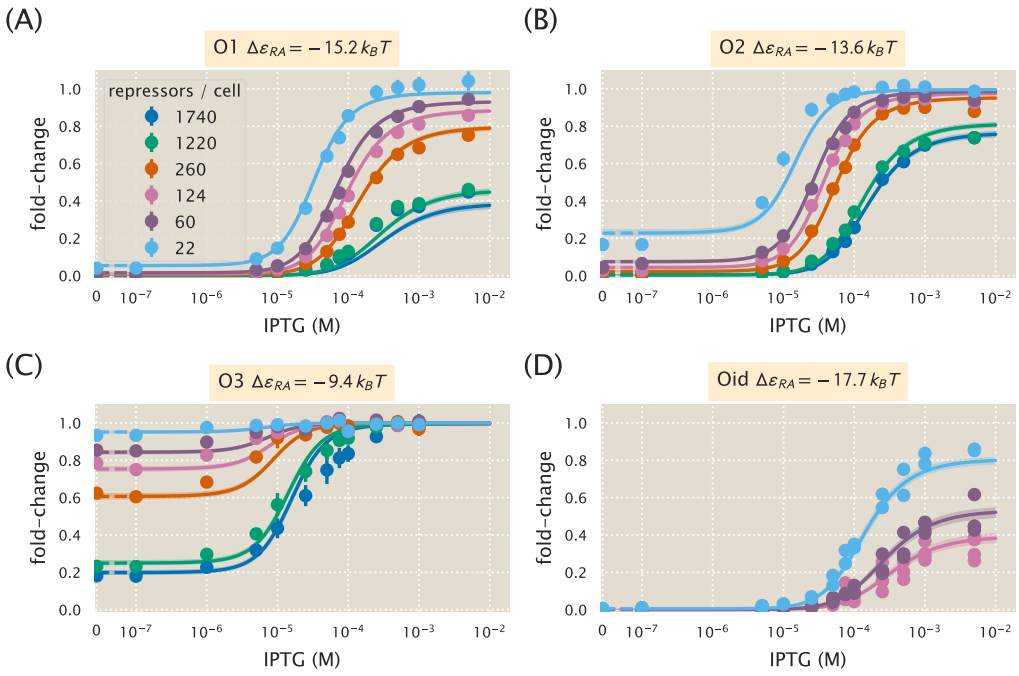

Figure 2.26: **Global fit of dissociation constants, repressor copy numbers and binding energies.** Theoretical predictions resulting from simultaneously fitting the dissociation constants K_A and K_I, the six repressor copy numbers R, and the four repressor-DNA binding energies $\Delta\varepsilon_{RA}$ using the entire data set from Figure 2.5 as well as the microscopy data for the Oid operator. Error bars of experimental data show the standard error of the mean (eight or more replicates) and shaded regions denote the 95% credible region. Where error bars are not visible, they are smaller than the point itself. For the Oid operator, all of the data points are shown since a smaller number of replicates were taken. The shaded regions are significantly smaller than in Figure 2.5 because this fit was based on all data points, and hence the fit parameters are much more tightly constrained. The dashed lines at 0 IPTG indicate a linear scale, whereas solid lines represent a log scale.

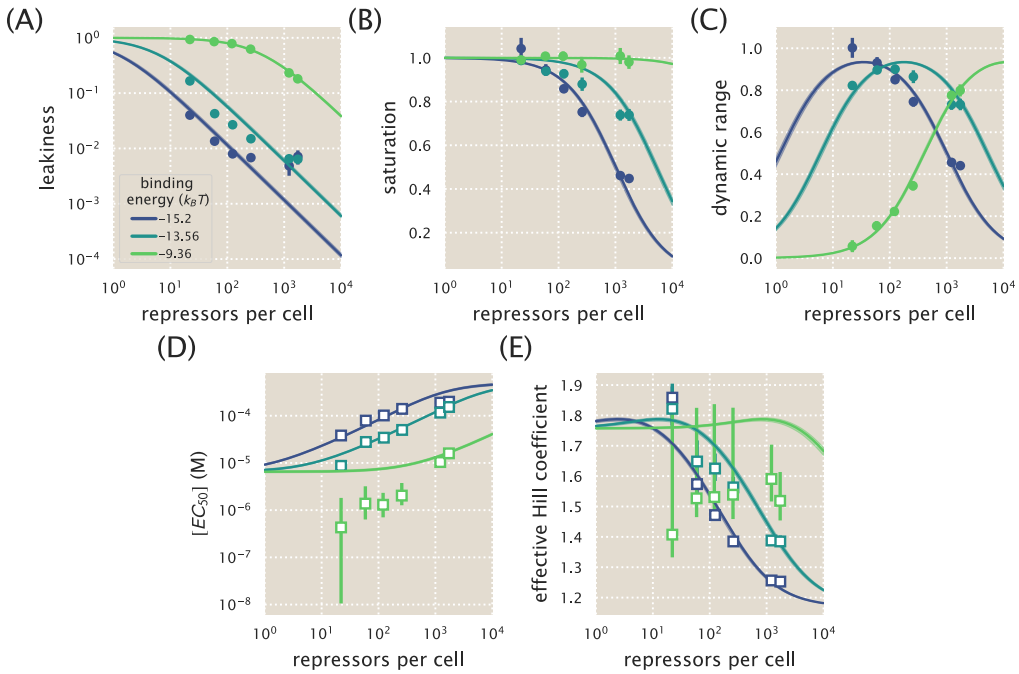

Figure 2.27: **Key properties of induction profiles as predicted with a global fit using all available data.** Data for the (A) leakiness, (B) saturation, and (C) dynamic range are obtained from fold-change measurements in Figure 2.5 in the absence and presence of IPTG. All prediction curves were generated using the parameters listed in 2.3. Both the (D) $[EC_{50}]$ and (E) effective Hill coefficient are inferred by individually fitting all parameters—K_A, K_I, R, $\Delta\varepsilon_{RA}$—to each operator-repressor pairing in Figure 2.5A-C separately to Equation 2.5 in order to smoothly interpolate between the data points. Note that where error bars are not visible, this indicates that the error bars are smaller than the point itself.

Table 2.3: **Global fit of all parameter values using the entire data set in Figure 2.5.** In addition to fitting the repressor inducer dissociation constants K_A and K_I as was done in the text, we also fit the repressor DNA binding energy $\Delta\varepsilon_{RA}$ as well as the repressor copy numbers R for each strain. The middle columns show the previously reported values for all $\Delta\varepsilon_{RA}$ and R values, with ± representing the standard deviation of three replicates. The right column shows the global fits from this work, with the subscript and superscript notation denoting the 95% credible region. Note that there is overlap between all of the repressor copy numbers and that the net difference in the repressor-DNA binding energies is less than 1 k_BT. The logarithms $\tilde{k}_A = -\log \frac{K_A}{1\,\mathrm{M}}$ and $\tilde{k}_I = -\log \frac{K_I}{1\,\mathrm{M}}$ of the dissociation constants were fit for numerical stability.

	Reported Values [9]	**Global Fit**
$\tilde{k}_A$	–	$-5.33^{+0.06}_{-0.05}$
$\tilde{k}_I$	–	$0.31^{+0.05}_{-0.06}$
K_A	–	$205^{+11}_{-12}\ \mu$M
K_I	–	$0.73^{+0.04}_{-0.04}\ \mu$M
R_{22}	22 ± 4	20^{+1}_{-1}
R_{60}	60 ± 20	74^{+4}_{-3}
R_{124}	124 ± 30	130^{+6}_{-6}
R_{260}	260 ± 40	257^{+9}_{-11}
R_{1220}	1220 ± 160	1191^{+32}_{-55}
R_{1740}	1740 ± 340	1599^{+75}_{-87}
O1 $\Delta\varepsilon_{RA}$	$-15.3 \pm 0.2\ k_BT$	$-15.2^{+0.1}_{-0.1}\ k_BT$
O2 $\Delta\varepsilon_{RA}$	$-13.9 \pm 0.2\ k_BT$	$-13.6^{+0.1}_{-0.1}\ k_BT$
O3 $\Delta\varepsilon_{RA}$	$-9.7 \pm 0.1\ k_BT$	$-9.4^{+0.1}_{-0.1}\ k_BT$
Oid $\Delta\varepsilon_{RA}$	$-17.0 \pm 0.2\ k_BT$	$-17.7^{+0.2}_{-0.1}\ k_BT$

2.12 Supplemental Information: Applicability of Theory to the Oid Operator Sequence

In addition to the native operator sequences (O1, O2, and O3) considered in the main text, we were also interested in testing our model predictions against the synthetic Oid operator. In contrast to the other operators, Oid is one base pair shorter in length (20 bp), is fully symmetric, and is known to provide stronger repression than the native operator sequences considered so far. While the theory should be similarly applicable, measuring the lower fold-changes associated with this YFP construct was expected to be near the sensitivity limit for our flow cytometer, due to the especially strong binding energy of Oid ($\Delta\varepsilon_{RA} = -17.0\ k_BT$) [13]. Accordingly, fluorescence data for Oid were obtained using microscopy, which is more sensitive than flow cytometry. Supplemental Section 2.8 gives a detailed explanation of how microscopy measurements were used to obtain induction curves.

We follow the approach of the main text and make fold-change predictions based on the parameter estimates from our strain with $R = 260$ and an O2 operator. These predictions are shown in Figure 2.28A, where we also plot data taken in triplicate for strains containing $R = 22$, 60, and 124, obtained by single-cell microscopy. We find that the data are systematically below the theoretical predictions. We also considered our global fitting approach (see Supplemental Section 2.11) to see whether we might find better agreement with the observed data. Interestingly, we findthat the majority of the parameters remain largely unchanged, but our estimate for the Oid binding energy $\Delta\varepsilon_{RA}$ is shifted to $-17.7\ k_BT$ instead of the value $-17.0\ k_BT$ found by Ref. [9]. In Figure 2.28B we again plot the Oid fold-change data but with theoretical predictions using the new estimate for the Oid binding energy from our global fit and find substantially better agreement.

Figure 2.29 shows the cumulative data from Ref. [9] and Ref. [10], as well as our data with $c = 0\ \mu$M, which all measured fold-change for the same simple repression architecture utilizing different reporters and measurement techniques. We find that the binding energies from the global fit, including $\Delta\varepsilon_{RA} = -17.7\ k_BT$, compare reasonably well with all previous measurements.

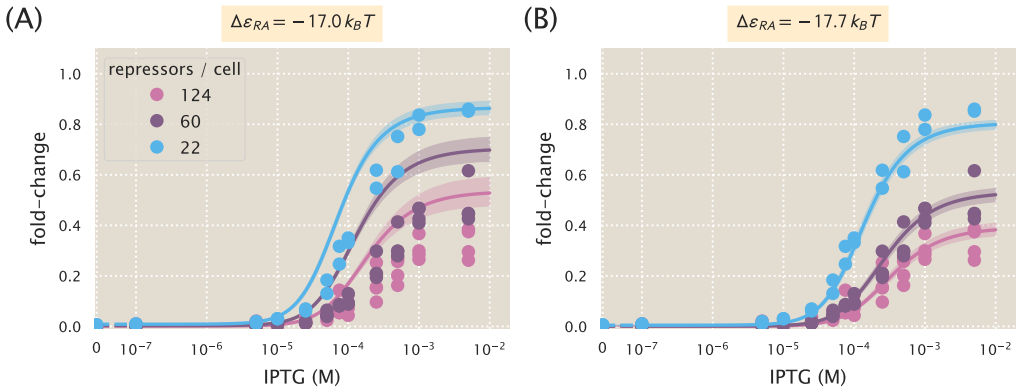

Figure 2.28: **Predictions of fold-change for strains with an Oid binding sequence versus experimental measurements with different repressor copy numbers.** (A) Experimental data is plotted against the parameter-free predictions that are based on our fit to the O2 strain with $R = 260$. Here we use the previously measured binding energy $\Delta\varepsilon_{RA} = -17.0\ k_BT$ [9]. (B) The same experimental data is plotted against the best-fit parameters using the complete O1, O2, O3, and Oid data sets to infer K_A, K_I, repressor copy numbers, and the binding energies of all operators (see Supplemental Section 2.11). Here the major difference in the inferred parameters is a shift in the binding energy for Oid from $\Delta\varepsilon_{RA} = -17.0\ k_BT$ to $\Delta\varepsilon_{RA} = -17.7\ k_BT$, which now shows agreement between the theoretical predictions and experimental data. Shaded regions from the theoretical curves denote the 95% credible region. These are narrower in Panel B because the inference of parameters was performed with much more data, and hence the best-fit values are more tightly constrained. Individual data points are shown due to the small number of replicates. The dashed lines at 0 IPTG indicate a linear scale, whereas solid lines represent a log scale.

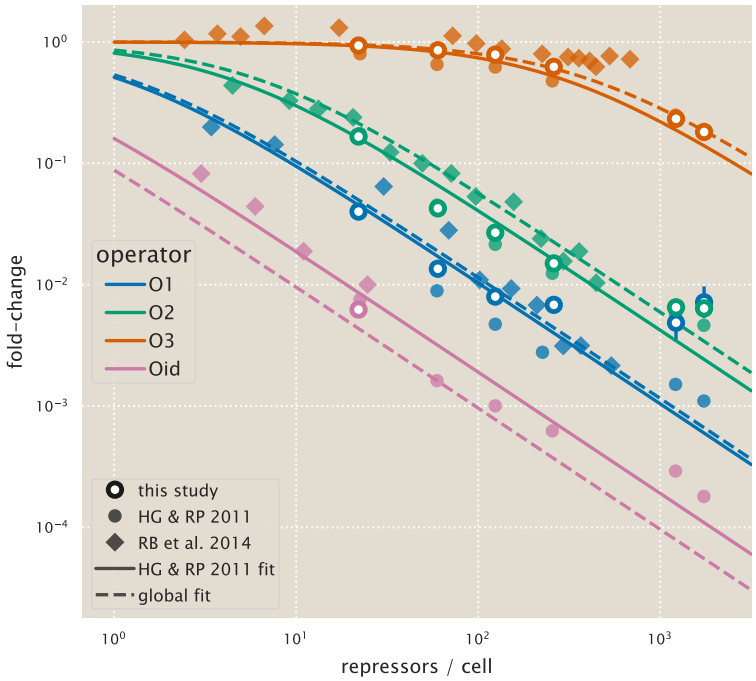

Figure 2.29: **Comparison of fold-change predictions based on binding energies from Garcia and Phillips and those inferred from this work.** Fold-change curves for the different repressor-DNA binding energies $\Delta\varepsilon_{RA}$ are plotted as a function of repressor copy number when IPTG concentration $c = 0$. Solid curves use the binding energies determined from Ref. [9], while the dashed curves use the inferred binding energies we obtained when performing a global fit of K_A, K_I, repressor copy numbers, and the binding energies using all available data from our work. Fold-change measurements from our experiments (outlined circles) Ref. [9] (solid circles), and Ref. [10] (diamonds) show that the small shifts in binding energy that we infer are still in agreement with prior data. Note that only a single flow cytometry data point is shown for Oid from this study, since the $R = 60$ and $R = 124$ curves from Figure 2.28 had extremely low fold-change in the absence of inducer ($c = 0$) so as to be indistinguishable from autofluorescence, and in fact their fold-change values in this limit were negative and hence do not appear on this plot.

2.13 Supplemental Information: Comparison of Parameter Estimation and Fold-Change Predictions across Strains

The inferred parameter values for K_A and K_I in the main text were determined by fitting to induction fold-change measurements from a single strain ($R = 260$, $\Delta\varepsilon_{RA} = -13.9\ k_BT$, $n = 2$, and $\Delta\varepsilon_{AI} = 4.5\ k_BT$). After determining these parameters, we were able to predict the fold-change of the remaining strains without any additional fitting. However, the theory should be independent of the specific strain used to estimate K_A and K_I; using any alternative strain to fit K_A and K_I should yield similar predictions. For the sake of completeness, here we discuss the values for K_A and K_I that are obtained by fitting to each of the induction data sets individually. These fit parameters are shown in Figure 2.5D of the main text, where we find close agreement between strains, but with some deviation and poorer inferences observed with the O3 operator strains. Overall, we find that regardless of which strain is chosen to determine the unknown parameters, the predictions laid out by the theory closely match the experimental measurements. Here we present a comparison of the strain specific predictions and measured fold-change data for each of the three operators considered.

We follow the approach taken in the main text and use Equation 2.5 to infer values for K_A and K_I by fitting to each combination of binding energy $\Delta\varepsilon_{RA}$ and repressor copy number R. We then use these fitted parameters to predict the induction curves of all other strains. In Figure 2.30 we plot these fold-change predictions along with experimental data for each of our strains that contains an O1 operator. To make sense of this plot consider the first row as an example. In the first row, K_A and K_I were estimated using data from the strain containing $R = 22$ and an O1 operator (top leftmost plot, shaded in gray). The remaining plots in this row show the predicted fold-change using these values for K_A and K_I. In each row, we then infer K_A and K_I using data from a strain containing a different repressor copy number ($R = 60$ in the second row, $R = 124$ in the third row, and so on). In Figure 2.31 and Figure 2.32, we similarly apply this inference to our strains with O2 and O3 operators, respectively. We note that the overwhelming majority of predictions closely match the experimental data.The notable exception is that using the $R = 22$ strain provides poor predictions for the strains with large copy numbers (especially $R = 1220$ and $R = 1740$), though it should be noted that predictions made from the $R = 22$ strain have considerably broader credible regions. This loss in predictive power is due to the poorer estimates of K_A and K_I for the $R = 22$ strain as shown in Figure 2.5D.

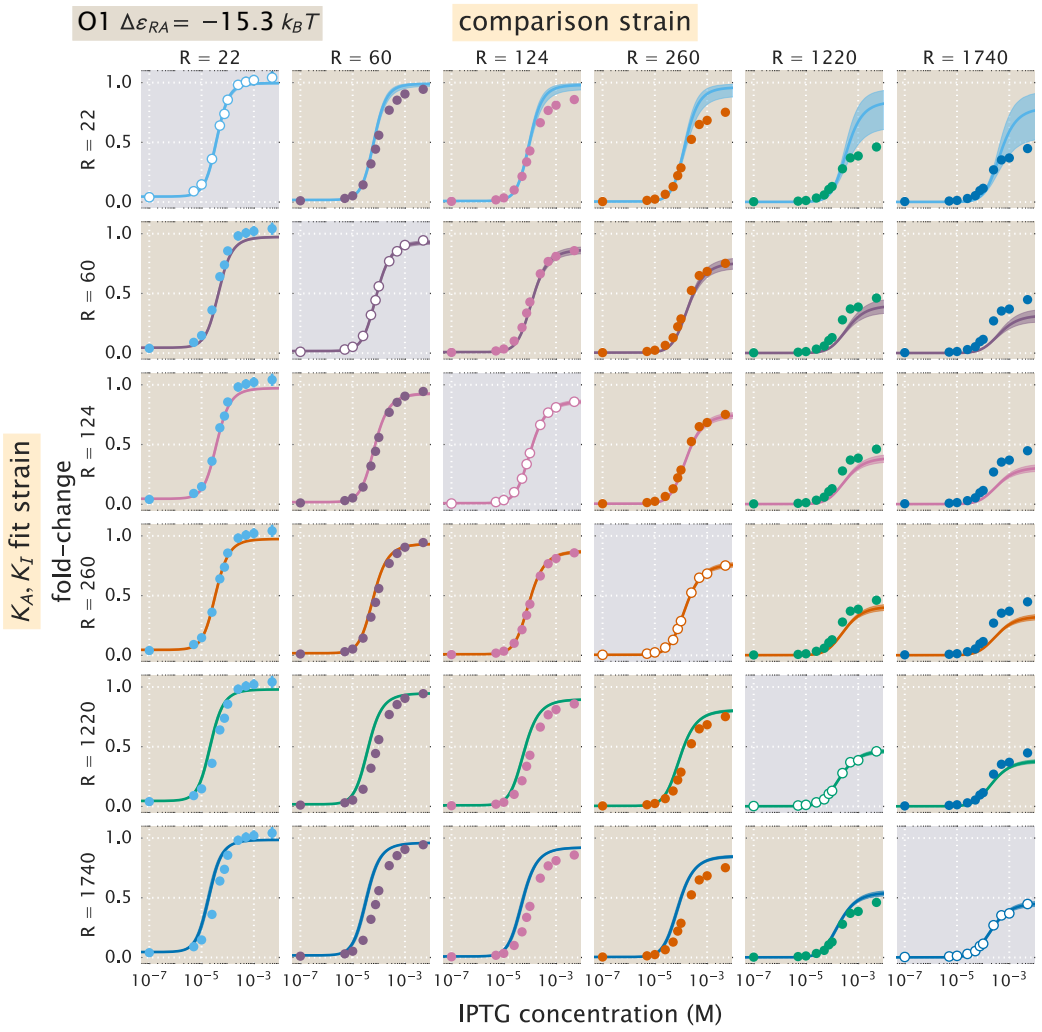

Figure 2.30: **O1 strain fold-change predictions based on strain-specific parameter estimation of K_A and K_I.** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O1 operator. The solid points correspond to the mean experimental value. The solid lines correspond to Equation 2.5 using the parameter estimates of K_A and K_I. Each row uses a single set of parameter values based on the strain noted on the left axis. The shaded plots along the diagonal are those where the parameter estimates are plotted along with the data used to infer them. Values for repressor copy number and operator binding energy are from Ref. [9]. The shaded region on the curve represents the uncertainty from our parameter estimates and reflects the 95% highest probability density region of the parameter predictions.

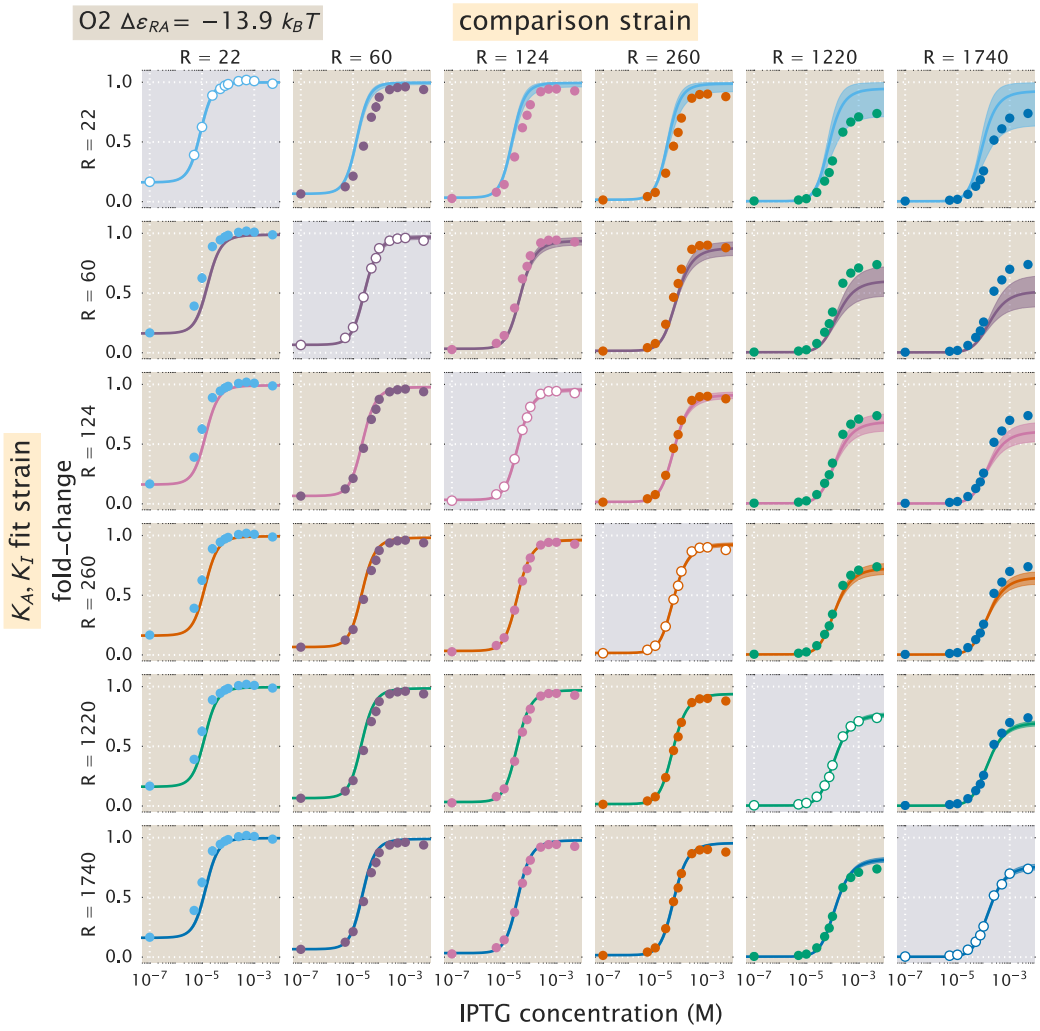

Figure 2.31: **O2 strain fold-change predictions based on strain-specific parameter estimation of K_A and K_I.** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O2 operator. The plots and data shown are analogous to Figure 2.30, but for the O2 operator.

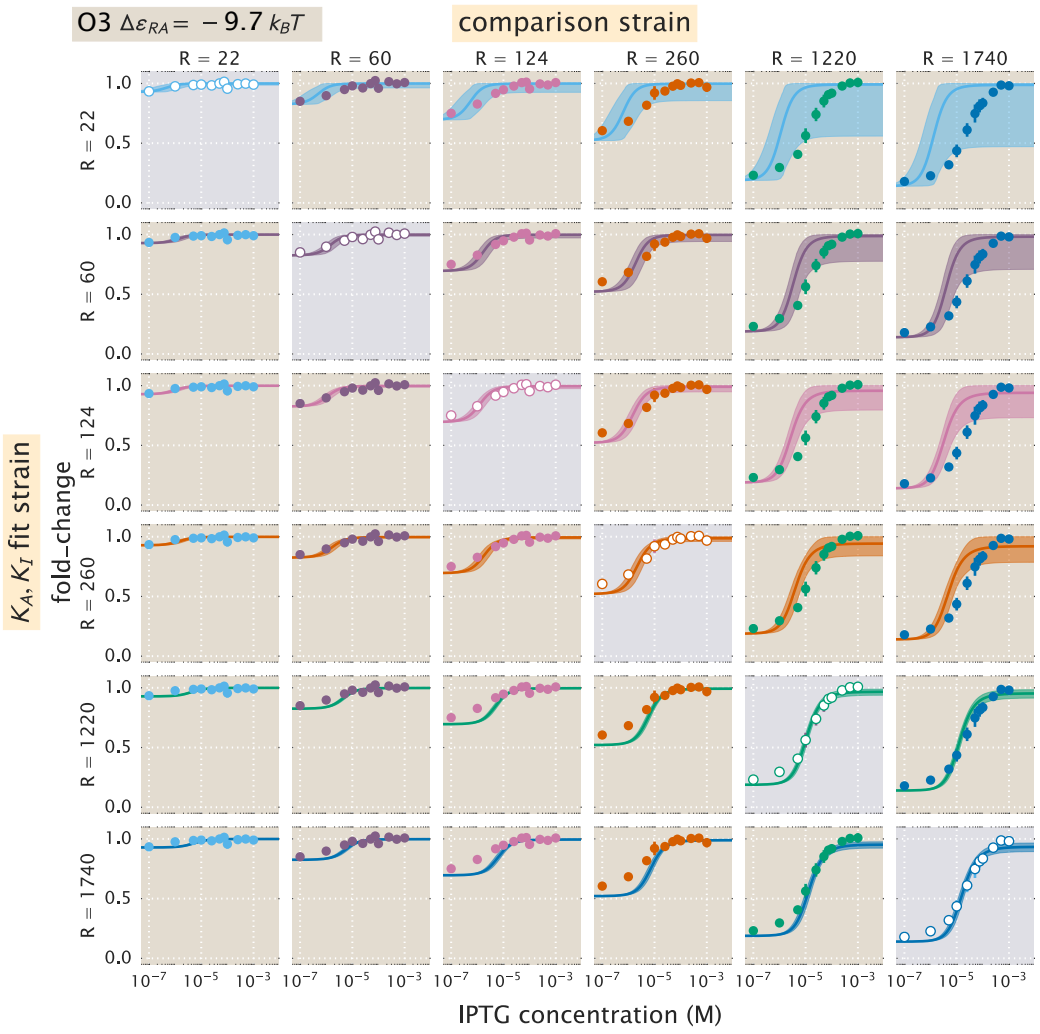

Figure 2.32: **O3 strain fold-change predictions based on strain-specific parameter estimation of K_A and K_I.** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O3 operator. The plots and data shown are analogous to Figure 2.30, but for the O3 operator. We note that when using the $R = 22$ O3 strain to predict K_A and K_I, the large uncertainty in the estimates of these parameters (see Figure 2.5D) leads to correspondingly wider credible regions.

2.14 Supplemental Information: Properties of Induction Titration Curves

In this section, we expand on the phenotypic properties of the induction response that were explored in the main text (see Figure 2.1). We begin by expanding on our discussion of dynamic range and then show the analytic form of the $[EC_{50}]$ for simple repression.

As stated in the main text, the dynamic range is defined as the difference between the maximum and minimum system response, or equivalently, as the difference between the saturation and leakiness of the system. Using Equations 2.6, 2.7, and 2.8, the dynamic range is given by

$$\text{dynamic range} = \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1} - \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}. \tag{2.52}$$

The dynamic range, saturation, and leakiness were plotted with our experimental data in Figure 2.6A-C as a function of repressor copy number. Figure 2.33 shows how these properties are expected to vary as a function of the repressor-operator binding energy. Note that the resulting curves for all three properties have the same shape as in Figure 2.6A-C, since the dependence of the fold-change upon the repressor copy number and repressor-operator binding energy are both contained in a single multiplicative term, $Re^{-\beta\Delta\varepsilon_{RA}}$. Hence, increasing R on a logarithmic scale (as in Figure 2.6A-C) is equivalent to decreasing $\Delta\varepsilon_{RA}$ on a linear scale (as in Figure 2.33).

An interesting aspect of the dynamic range is that it exhibits a peak as a function of either the repressor copy number (or equivalently of the repressor-operator binding energy). Differentiating the dynamic range Equation 2.52 and setting it equal to zero, we find that this peak occurs at

$$\frac{R^*}{N_{NS}} = e^{-\beta(\Delta\varepsilon_{AI} - \Delta\varepsilon_{RA})}\sqrt{e^{\Delta\varepsilon_{AI}} + 1}\sqrt{e^{\Delta\varepsilon_{AI}} + \left(\frac{K_A}{K_I}\right)^n}. \tag{2.53}$$

The magnitude of the peak is given by

$$\text{max dynamic range} = \frac{\left(\sqrt{e^{\Delta\varepsilon_{AI}} + 1} - \sqrt{e^{\Delta\varepsilon_{AI}} + \left(\frac{K_A}{K_I}\right)^n}\right)^2}{\left(\frac{K_A}{K_I}\right)^n - 1}, \tag{2.54}$$

which is independent of the repressor-operator binding energy $\Delta\varepsilon_{RA}$ or R, and will only cause a shift in the location of the peak but not its magnitude.

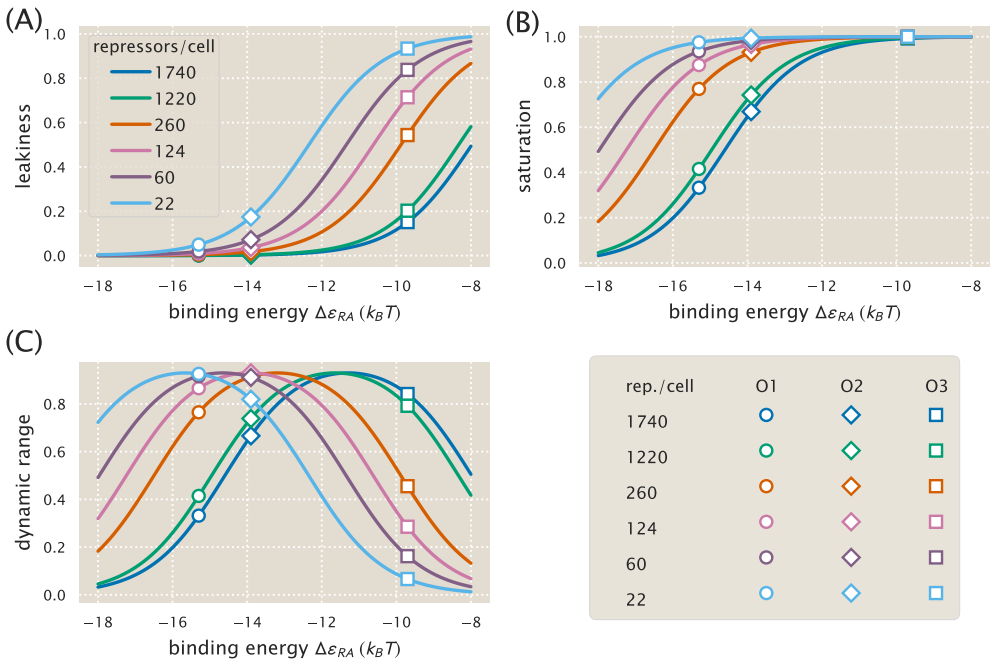

Figure 2.33: **Dependence of leakiness, saturation, and dynamic range on the operator binding energy and repressor copy number.** Increasing repressor copy number or decreasing the repressor-operator binding energy suppresses gene expression and decreases both the (A) leakiness and (B) saturation. (C) The dynamic range retains its shape but shifts right as the repressor copy number increases. The peak in the dynamic range can be understood by considering the two extremes for $\Delta\varepsilon_{RA}$: for small repressor-operator binding energies, the leakiness is small but the saturation increases with $\Delta\varepsilon_{RA}$; for large repressor-operator binding energies the saturation is near unity and the leakiness increases with $\Delta\varepsilon_{RA}$, thereby decreasing the dynamic range. Repressor copy number does not affect the maximum dynamic range (see Equation 2.54). Circles, diamonds, and squares represent $\Delta\varepsilon_{RA}$ values for the O1, O2, and O3 operators, respectively, demonstrating the expected values of the properties using those strains.

We now consider the two remaining properties, the $[EC_{50}]$ and effective Hill coefficient, which determine the horizontal properties of a system. That is, they determine the range of inducer concentration in which the system's response goes from its minimum to maximum values. The $[EC_{50}]$ denotes the inducer concentration required to generate fold-change halfway between its minimum and maximum value and was defined implicitly in Equation 2.9. For the simple repression system, the $[EC_{50}]$ is

given by

$$\frac{[EC_{50}]}{K_A} = \frac{\frac{K_A}{K_I} - 1}{\frac{K_A}{K_I} - \left(\frac{\left(1+\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)+\left(\frac{K_A}{K_I}\right)^n\left(2e^{-\beta\Delta\varepsilon_{AI}}+\left(1+\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)\right)}{2\left(1+\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)+e^{-\beta\Delta\varepsilon_{AI}}+\left(\frac{K_A}{K_I}\right)^n e^{-\beta\Delta\varepsilon_{AI}}} \right)^{\frac{1}{n}}} - 1. \quad (2.55)$$

Using this expression, we can then find the effective Hill coefficient h, which equals twice the log-log slope of the normalized fold-change evaluated at $c = [EC_{50}]$ (see Equation 2.10). In Figure 2.6D-E we show how these two properties vary with repressor copy number, and in Figure 2.34 we demonstrate how they depend on the repressor-operator binding energy. Both the $[EC_{50}]$ and h vary significantly with repressor copy number for sufficiently strong operator binding energies. Interestingly, for weak operator binding energies on the order of the O3 operator, it is predicted that the effective Hill coefficient should not vary with repressor copy number. In addition, the maximum possible Hill coefficient is roughly 1.75, which stresses the point that the effective Hill coefficient should not be interpreted as the number of inducer binding sites, which is exactly 2.

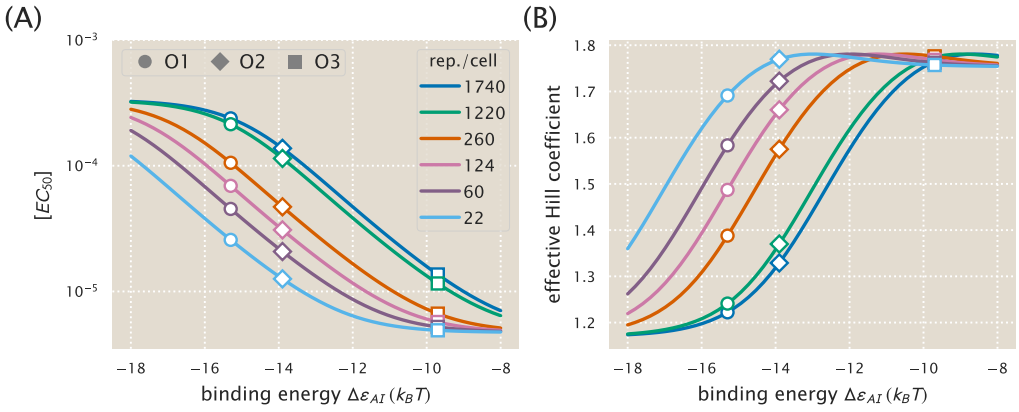

Figure 2.34: **$[EC_{50}]$ and effective Hill coefficient depend strongly on repressor copy number and operator binding energy.** (A) $[EC_{50}]$ values range from very small and tightly clustered at weak operator binding energies (e.g. O3) to relatively large and spread out for stronger operator binding energies (O1 and O2). (B) The effective Hill coefficient generally decreases with increasing repressor copy number, indicating a flatter normalized response. The maximum possible Hill coefficient is roughly 1.75 for all repressor-operator binding energies. Circles, diamonds, and squares represent $\Delta\varepsilon_{RA}$ values for the O1, O2, and O3 operators, respectively.

2.15 Supplemental Information: Applications to Other Regulatory Architectures

In this section, we discuss how the theoretical framework presented in this work is sufficiently general to include a variety of regulatory architectures outside of simple repression by LacI. We begin by noting that the exact same formula for fold-change given in Equation 2.5 can also describe corepression. We then demonstrate how our model can be generalized to include other architectures, such as a coactivator binding to an activator to promote gene expression. In each case, we briefly describe the system and describe its corresponding theoretical description. For further details, we invite the interested reader to read Refs. [24, 28].

Corepression

Consider a regulatory architecture where binding of a transcriptional repressor occludes the binding of RNAP to the DNA. A corepressor molecule binds to the repressor and shifts its allosteric equilibrium towards the active state in which it binds more tightly to the DNA, thereby decreasing gene expression (in contrast, an inducer shifts the allosteric equilibrium towards the inactive state where the repressor binds more weakly to the DNA). As in the main text, we can enumerate the states and statistical weights of the promoter and the allosteric states of the repressor. We note that these states and weights exactly match Figure 2.2 and yield the same fold-change equation as Equation 2.5,

$$\text{fold-change} \approx \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1 + \frac{c}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{2.56}$$

where c now represents the concentration of the corepressor molecule. Mathematically, the difference between these two architectures can be seen in the relative sizes of the dissociation constants K_A and K_I between the inducer and repressor in the active and inactive states, respectively. The corepressor is defined by $K_A < K_I$, since the corepressor favors binding to the repressor's active state; an inducer must satisfy $K_I < K_A$, as was found in the main text from the induction data (see Figure 2.4). Much as was performed in the main text, we can make some predictions about the how the response of a corepressor. In Figure 2.35A, we show how varying the repressor copy number R and the repressor-DNA binding energy $\Delta\varepsilon_{RA}$ influences the response. We draw the reader's attention to the decrease in fold-change as the concentration of effector is increased.

Activation

We now turn to the case of activation. While this architecture was not studied in this work, we wish to demonstrate how the framework presented here can be extended to include transcription factors other than repressors. To that end, we consider a transcriptional activator which binds to DNA and aids in the binding of RNAP through the energetic interaction term ε_{AP}. Note that in this architecture, binding of the activator does not occlude binding of the polymerase. Binding of a coactivator molecule shifts its allosteric equilibrium towards the active state ($K_A < K_I$), where the activator is more likely to bind to the DNA and promote expression. Enumerating all of the states and statistical weights of this architecture and making the approximation that the promoter is weak generates a fold-change equation of the form

$$\text{fold-change} = \frac{1 + \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n} \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_{AA}} e^{-\beta\varepsilon_{AP}}}{1 + \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n} \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_{AA}}}, \tag{2.57}$$

where A is the total number of activators per cell, c is the concentration of a coactivator molecule, $\Delta\varepsilon_{AA}$ is the binding energy of the activator to the DNA in the active allosteric state, and ε_{AP} is the interaction energy between the activator and the RNAP. Unlike in the cases of induction and corepression, the fold-change formula for activation includes terms from when the RNAP is bound by itself on the DNA as well as when both RNAP and the activator are simultaneously bound to the DNA. Figure 2.35B explores predictions of the fold-change in gene expression by manipulating the activator copy number, DNA binding energy, and the polymerase-activator interaction energy. Note that with this activation scheme, the fold-change must necessarily be greater than one. An interesting feature of these predictions is the observation that even small changes in the interaction energy ($< 0.5\ k_BT$) can result in dramatic increase in fold-change.

As in the case of induction, the Equation 2.57 is straightforward to generalize. For example, the relative values of K_I and K_A can be switched such that $K_I < K_A$ in which the secondary molecule drives the activator to assume the inactive state represents induction of an activator. While these cases might be viewed as separate biological phenomena, mathematically they can all be described by the same underlying formalism.

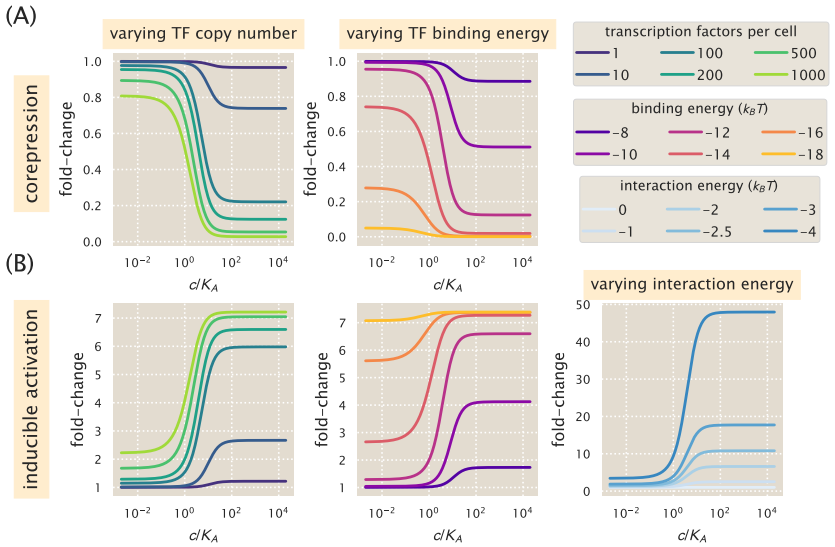

Figure 2.35: **Representative fold-change predictions for allosteric corepression and activation.** (A) Contrary to the case of induction described in the main text, addition of a corepressor decreases fold-change in gene expression. The left and right panels demonstrate how varying the values of the repressor copy number R and repressor-DNA binding energy $\Delta\varepsilon_{RA}$, respectively, change the predicted response profiles. (B) In the case of inducible activation, binding of an effector molecule to an activator transcription factor increases the fold-change in gene expression. Note that for activation, the fold-change is greater than 1. The left and center panels show how changing the activator copy number A and activator-DNA binding energy $\Delta\varepsilon_{AA}$ alter response, respectively. The right panel shows how varying the polymerase-activator interaction energy ε_{AP} alters the fold-change. Relatively small perturbations to this energetic parameter drastically change the level of activation and play a major role in dictating the dynamic range of the system.

2.16 Supplemental Information: *E. coli* Primer and Strain List

Here we provide additional details about the genotypes of the strains used, as well as the primer sequences used to generate them. *E. coli* strains were derived from K12 MG1655. For those containing $R = 22$, we used strain HG104 which additionally has the *lacYZA* operon deleted (positions 360,483 to 365,579) but still contains the native *lacI* locus. All other strains used strain HG105, where both the *lacYZA* and *lacI* operons have both been deleted (positions 360,483 to 366,637).

All 25x+11-yfp expression constructs were integrated at the *galK* locus (between positions 1,504,078 and 1,505,112) while the 3*1x-lacI constructs were integrated at the *ybcN* locus (between positions 1,287,628 and 1,288,047). Integration was performed with λ Red recombineering [63] as described in Ref. [9] using the primers listed in Table 4.2. We follow the notation of Lutz and Bujard [53] for the nomenclature of the different constructs used. Specifically, the first number refers to the antibiotic resistance cassette that is present for selection (2 = kanamycin, 3 = chloramphenicol, and 4 = spectinomycin) and the second number refers to the promoter used to drive expression of either YFP or LacI (1 = $P_{LtetO-1}$, and 5 = *lacUV5*). Note that in 25x+11-yfp, x refers to the LacI operator used, which is centered at +11 (or alternatively, begins at the transcription start site). For the different LacI constructs, 3*1x-lacI, x refers to the different ribosomal binding site modifications that provide different repressor copy numbers and follows from Ref. [9]. The asterisk refers to the presence of FLP recombinase sites flanking the chloramphenicol resistance gene that can be used to lose this resistance. However, we maintained the resistance gene in our constructs. A summary of the final genotypes of each strain is listed in Table 2.5. In addition each strain also contained the plasmid pZS4*1-mCherry and provided constitutive expression of the mCherry fluorescent protein. This pZS plasmid is a low copy (SC101 origin of replication) where like with 3*1x-lacI, mCherry is driven by a $P_{LtetO-1}$ promoter.

Table 2.4: **Promoter sequences and primers used in this work.** The listed promoter sequences were randomly mutated to produce libraries for use in Sort-Seq experiments. The primer sequences were used to generate plasmids for Sort-Seq experiments or for use in creating strains with mutated operators or LacI.

Primer	**Sequence**	**Comments**
General sequencing primers		
pZSforwseq2	TTCCCAACCTTACCAGAGGGC	Forward primer for 3*1x-lacI
251F	CCTTTCGTCTTCACCTCGA	Forward primer for 25x+11-yfp
YFP1	ACTAGCAACACCAGAACAGCCC	Reverse primer for 3*1x-lacI and 25x+11-yfp
Integration primers:		
HG6.1 (*galK*)	gtttgcgcgcagtcagcgatatccattttcgcgaatccgg agtgtaagaaACTAGCAACACCAGAACAGCC	Reverse primer for 25x+11-yfp with homology to *galK* locus.
HG6.3 (*galK*)	ttcatattgttcagcgacagcttgctgtacggcaggcacc agctcttccgGGCTAATGCACCCAGTAAGG	Forward primer for 25x+11-yfp with homology to *galK* locus.
galK-control-upstream1	TTCATATTGTTCAGCGACAGCTTG	To check integration.
galK-control-downstream1	CTCCGCCACCGTACGTAAATT	To check integration.
HG11.1 (*ybcN*)	acctctgcggaggggaagcgtgaacctctcacaagacggc atcaaattacACTAGCAACACCAGAACAGCC	Reverse primer for 3*1x-lacI with homology to *ybcN* locus.
HG11.3 (*ybcN*)	ctgtagatgtgtccgttcatgacacgaataagcggtgtag ccattacgccGGCTAATGCACCCAGTAAGG	Forward primer for 3*1x-lacI with homology to *ybcN* locus.
ybcN-control-upstream1	AGCGTTTGACCTCTGCGGA	To check integration.
ybcN-control-downstream1	GCTCAGGTTTACGCTTACGACG	To check integration.

Table 2.5: ***E. coli* strains used in this work.** Each strain contains a unique operator-yfp construct for measurement of fluorescence and R refers to the dimer copy number as measured by Ref. [9].

Strain	**Genotype**
O1, $R = 0$	HG105::galK⟨⟩25O1+11-yfp
O1, $R = 22$	HG104::galK⟨⟩25O1+11-yfp
O1, $R = 60$	HG105::galK⟨⟩25O1+11-yfp, ybcN⟨⟩3*1RBS1147-lacI
O1, $R = 124$	HG105::galK⟨⟩25O1+11-yfp, ybcN⟨⟩3*1RBS1027-lacI
O1, $R = 260$	HG105::galK⟨⟩25O1+11-yfp, ybcN⟨⟩3*1RBS446-lacI
O1, $R = 1220$	HG105::galK⟨⟩25O1+11-yfp, ybcN⟨⟩3*1RBS1-lacI
O1, $R = 1740$	HG105::galK⟨⟩25O1+11-yfp, ybcN⟨⟩3*1-lacI (RBS1L)
O2, $R = 0$	HG105::galK⟨⟩25O2+11-yfp
O2, $R = 22$	HG104::galK⟨⟩25O2+11-yfp
O2, $R = 60$	HG105::galK⟨⟩25O2+11-yfp, ybcN⟨⟩3*1RBS1147-lacI
O2, $R = 124$	HG105::galK⟨⟩25O2+11-yfp, ybcN⟨⟩3*1RBS1027-lacI
O2, $R = 260$	HG105::galK⟨⟩25O2+11-yfp, ybcN⟨⟩3*1RBS446-lacI
O2, $R = 1220$	HG105::galK⟨⟩25O2+11-yfp, ybcN⟨⟩3*1RBS1-lacI
O2, $R = 1740$	HG105::galK⟨⟩25O2+11-yfp, ybcN⟨⟩3*1-lacI (RBS1L)
O3, $R = 0$	HG105::galK⟨⟩25O3+11-yfp
O3, $R = 22$	HG104::galK⟨⟩25O3+11-yfp
O3, $R = 60$	HG105::galK⟨⟩25O3+11-yfp, ybcN⟨⟩3*1RBS1147-lacI
O3, $R = 124$	HG105::galK⟨⟩25O3+11-yfp, ybcN⟨⟩3*1RBS1027-lacI
O3, $R = 260$	HG105::galK⟨⟩25O3+11-yfp, ybcN⟨⟩3*1RBS446-lacI
O3, $R = 1220$	HG105::galK⟨⟩25O3+11-yfp, ybcN⟨⟩3*1RBS1-lacI
O3, $R = 1740$	HG105::galK⟨⟩25O3+11-yfp, ybcN⟨⟩3*1-lacI (RBS1L)
Oid, $R = 0$	HG105::galK⟨⟩25Oid+11-yfp
Oid, $R = 22$	HG104::galK⟨⟩25Oid+11-yfp
Oid, $R = 60$	HG105::galK⟨⟩25Oid+11-yfp, ybcN⟨⟩3*1RBS1147-lacI
Oid, $R = 124$	HG105::galK⟨⟩25Oid+11-yfp, ybcN⟨⟩3*1RBS1027-lacI
Oid, $R = 260$	HG105::galK⟨⟩25Oid+11-yfp, ybcN⟨⟩3*1RBS446-lacI
Oid, $R = 1220$	HG105::galK⟨⟩25Oid+11-yfp, ybcN⟨⟩3*1RBS1-lacI
Oid, $R = 1740$	HG105::galK⟨⟩25Oid+11-yfp, ybcN⟨⟩3*1-lacI (RBS1L)

BIBLIOGRAPHY

[1] Janet E. Lindsley and Jared Rutter. Whence cometh the allosterome? *Proceedings of the National Academy of Sciences*, 103(28):10533–5, 2006.

[2] James G. Harman. Allosteric regulation of the cAMP receptor protein. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, 1547(1):1–17, 2001.

[3] Maria Fe Lanfranco, Fernanda Gárate, Ashton J. Engdahl, and Rodrigo A. Maillard. Asymmetric configurations in a reengineered homodimer reveal multiple subunit communication pathways in protein allostery. *The Journal of Biological Chemistry*, 292(15):6086–6093, 2017.

[4] Yaki Setty, Avraham E. Mayo, Michael G. Surette, and Uri Alon. Detailed map of a cis-regulatory input function. *Proceedings of the National Academy of Sciences*, 100(13):7702–7707, 2003.

[5] Frank J. Poelwijk, Marjon G. J. deVos, and Sander J. Tans. Tradeoffs and optimality in the evolution of gene regulation. *Cell*, 146(3):462–470, 2011.

[6] José M. G. Vilar and Leonor Saiz. Reliable prediction of complex phenotypes from a modular design in free energy space: An extensive exploration of the *lac* operon. *ACS Synthetic Biology*, 2(10):576–586, 2013.

[7] Jameson K. Rogers, Christopher D. Guzman, Noah D. Taylor, Srivatsan Raman, Kelley Anderson, and George M. Church. Synthetic biosensors for precise gene control and real-time monitoring of metabolites. *Nucleic Acids Research*, 43(15):7648–7659, 2015.

[8] Julia Rohlhill, Nicholas R. Sandoval, and Eleftherios T Papoutsakis. Sort-seq approach to engineering a formaldehyde-inducible promoter for dynamically regulated *Escherichia coli* growth on methanol. *ACS Synthetic Biology*, page Advance online publication, 2017.

[9] Hernan G. Garcia and Rob Phillips. Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences*, 108(29):12173–8, 2011a.

[10] Robert C. Brewster, Franz M. Weinert, Hernan G. Garcia, Dan Song, Mattias Rydenfelt, and Rob Phillips. The transcription factor titration effect dictates level of gene expression. *Cell*, 156(6):1312–1323, 2014.

[11] Franz M. Weinert, Robert C. Brewster, Mattias Rydenfelt, Rob Phillips, and Willem K. Kegel. Scaling of gene expression with transcription-factor fugacity. *Physical Review Letters*, 113(25):1–5, 2014.

[12] Jacques Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, 12:88–118, 1965.

[13] Hernan G. Garcia, Heun Jin Lee, James Q. Boedicker, and Rob Phillips. Comparison and calibration of different reporters for quantitative analysis of gene expression. *Biophysical Journal*, 101(3):535–544, 2011b.

[14] Robert C. Brewster, Daniel L. Jones, and Rob Phillips. Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Computational Biology*, 8(12):e1002811, 2012.

[15] James Q. Boedicker, Hernan G. Garcia, and Rob Phillips. Theoretical and experimental dissection of DNA loop-mediated repression. *Physical Review Letters*, 110(1):018101, 2013.

[16] James Q. Boedicker, Hernan G. Garcia, Stephanie Johnson, and Rob Phillips. Dna sequence-dependent mechanics and protein-assisted bending in repressor-mediated loop formation. *Physical Biology*, 10(6):066005, 2013.

[17] Zhimin Huang, Liang Zhu, Yan Cao, Geng Wu, Xinyi Liu, Yingyi Chen, Qi Wang, Ting Shi, Yaxue Zhao, Yuefei Wang, Weihua Li, Yixue Li, Haifeng Chen, Guoqiang Chen, and Jian Zhang. Asd: a comprehensive database of allosteric proteins and modulators. *Nucleic Acids Research*, 39:D663, 2011.

[18] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S. Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–635, 2014.

[19] Gary K. Ackers, Alexander D. Johnson, and Madeline A. Shea. Quantitative model for gene regulation by lambda phage repressor. *Proceedings of the National Academy of Sciences*, 79(4):1129–33, 1982.

[20] Nicolas E. Buchler, Ulrich Gerland, and Terence Hwa. On schemes of combinatorial transcription logic. *PNAS*, 100(9):5136–41, 2003.

[21] José M. G. Vilar and Stanislas Leibler. DNA looping and physical constraints on transcription regulation. *Journal of Molecular Biology*, 331(5):981–989, 2003.

[22] Lacramioara Bintu, Nicolas E. Buchler, Hernan G. Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, Thomas Kuhlman, and Rob Phillips. Transcriptional regulation by the numbers: applications. *Current Opinion in Genetics & Development*, 15(2):125–135, 2005.

[23] Rob Phillips. Napoleon is in equilibrium. *Annual Review of Condensed Matter Physics*, 6(1):85–111, 2015.

[24] Lacramioara Bintu, Nicolas E. Buchler, Hernan G. Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips. Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development*, 15(2):116–124, 2005.

[25] Thomas Kuhlman, Zhongge Zhang, Milton H. Saier, and Terence Hwa. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 104(14):6043–8, 2007.

[26] Robert Daber, Matthew A. Sochor, and Mitchell Lewis. Thermodynamic analysis of mutant *lac* repressors. *Journal of Molecular Biology*, 409(1):76–87, 2011.

[27] Stefan Klumpp and Terence Hwa. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proceedings of the National Academy of Sciences*, 105(51):20245–50, 2008.

[28] Sarah Marzen, Hernan G. Garcia, and Rob Phillips. Statistical mechanics of Monod-Wyman-Changeux (MWC) models. *Journal of Molecular Biology*, 425(9):1433–1460, 2013.

[29] Ronald B. O'Gorman, John M. Rosenberg, Olga B. Kallai, Richard E. Dickerson, Keichi Itakura, Arthur D. Riggs, and Kathleen Shive Matthews. Equilibrium binding of inducer to *lac* repressor-operator DNA complex. *Journal of Biological Chemistry*, 255(21):10107–10114, 1980.

[30] Kevin F. Murphy, Gábor Balázsi, and James J. Collins. Combinatorial promoter design for engineering noisy gene expression. *Proceedings of the National Academy of Sciences*, 104(31):12726–12731, 2007.

[31] Robert Daber, Kim Sharp, and Mitchell Lewis. One is not enough. *Journal of Molecular Biology*, 392(5):1133–1144, 2009.

[32] Kevin F. Murphy, Rhys M. Adams, Xiao Wang, Gábor Balázsi, and James J. Collins. Tuning and controlling gene expression noise in synthetic gene networks. *Nucleic Acids Research*, 38(8):2712–2726, 2010.

[33] Matthew Almond Sochor. *In vitro* transcription accurately predicts *lac* repressor phenotype *in vivo* in *Escherichia coli*. *PeerJ*, 2:e498, 2014.

[34] Mattias Rydenfelt, Robert Sidney Cox, Hernan Garcia, and Rob Phillips. Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration. *Physical Review E*, 89:012702, 2014.

[35] Devinderjit Sivia and John Skilling. *Data analysis: a Bayesian tutorial*. OUP Oxford, 2006.

[36] Stefan Oehler, Michèle Amouyal, Peter Kolkhof, Brigitte von Wilcken-Bergmann, and Benno Müller-Hill. Quality and position of the three *lac* operators of *E. coli* define efficiency of repression. *The EMBO Journal*, 13(14):3348–3355, 1994.

[37] Matthew Scott, Carl W. Gunderson, Eduard M. Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of cell growth and gene expression: Origins and consequences. *Science*, 330(6007):1099–102, 2010.

[38] Jennifer A. N. Brophy and Christopher A. Voigt. Principles of genetic circuit design. *Nature Methods*, 11(5):508–520, 2014.

[39] David L. Shis, Faiza Hussain, Sarah Meinhardt, Liskin Swint-Kruse, and Matthew R. Bennett. Modular, multi-input transcriptional logic gating with orthogonal LacI/GalR family chimeras. *ACS Synthetic Biology*, 3(9):645–651, 2014.

[40] Bruno M. C. Martins and Peter S. Swain. Trade-Offs and constraints in allosteric sensing. *PLoS Computational Biology*, 7(11):1–13, 2011.

[41] Victor Sourjik and Howard C. Berg. Receptor sensitivity in bacterial chemotaxis. *Proceedings of the National Academy of Sciences*, 99(1):123–127, 2002.

[42] Juan E. Keymer, Robert G. Endres, Monica Skoge, Yigal Meir, and Ned S. Wingreen. Chemosensing in *Escherichia coli*: Two regimes of two-state receptors. *Proceedings of the National Academy of Sciences*, 103(6):1786–91, 2006.

[43] Lee R. Swem, Danielle L. Swem, Ned S. Wingreen, and Bonnie L. Bassler. Deducing receptor signaling parameters from *in vivo* analysis: LuxN/AI-1 quorum sensing in *Vibrio harveyi*. *Cell*, 134(3):461–473, 2008.

[44] Leonid A. Mirny. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences*, 107(52):22534–9, 2010.

[45] Tal Einav, Linas Mazutis, and Rob Phillips. Statistical mechanics of allosteric enzymes. *The Journal of Physical Chemistry B*, 121, 2016.

[46] Hernan G. Garcia, Alvaro Sanchez, James Q. Boedicker, Melisa Osborne, Jeff Gelles, Jane Kondev, and Rob Phillips. Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. *Cell Reports*, 2(1):150–161, 2012.

[47] Jacques Monod, Jean-Pierre Changeux, and François Jacob. Allosteric proteins and cellular control systems. *Journal of Molecular Biology*, 6:306–329, 1963.

[48] Anthony Auerbach. Thinking in cycles: MWC is a good model for acetylcholine receptor-channels. *The Journal of Physiology*, 590(1):93–8, 2012.

[49] Algirdas Velyvis, Ying R. Yang, Howard K. Schachman, and Lewis E. Kay. A solution NMR study showing that active site ligands and nucleotides directly perturb the allosteric equilibrium in aspartate transcarbamoylase. *Proceedings of the National Academy of Sciences*, 104(21):8815–20, 2007.

[50] Meritxell Canals, J. Robert Lane, Adriel Wen, Peter J. Scammells, Patrick M. Sexton, and Arthur Christopoulos. A Monod-Wyman-Changeux mechanism can explain G protein-coupled receptor (GPCR) allosteric modulation. *Journal of Biological Chemistry*, 287(1):650–659, 2012.

[51] Ron Milo, Jennifer H. Hou, Michael Springer, Michael P. Brenner, and Marc W. Kirschner. The relationship between evolutionary and physiological variation in hemoglobin. *Proceedings of the National Academy of Sciences*, 104(43):16998–17003, 2007.

[52] Matteo Levantino, Alessandro Spilotros, Marco Cammarata, Giorgio Schirò, Chiara Ardiccioni, Beatrice Vallone, Maurizio Brunori, and Antonio Cupane. The Monod-Wyman-Changeux allosteric model accounts for the quaternary transition dynamics in wild type and a recombinant mutant human hemoglobin. *Proceedings of the National Academy of Sciences*, 109(37):14894–9, 2012.

[53] Rolf Lutz and Hermann Bujard. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research*, 25(6):1203–10, 1997.

[54] Tae Seok Moon, Chunbo Lou, Alvin Tamsir, Brynne C. Stanton, and Christopher A. Voigt. Genetic programs constructed from layered logic gates in single cells. *Nature*, 491(7423):249–253, 2012.

[55] Leonor Saiz and Jose M. G. Vilar. Ab initio thermodynamic modeling of distal multisite transcription regulation. *Nucleic Acids Research*, 36(3):726, 2008.

[56] Sudheer Tungtur, Harlyn Skinner, Hongli Zhan, Liskin Swint-Kruse, and Dorothy Beckett. *In vivo* tests of thermodynamic models of transcription repressor function. *Biophysical Chemistry*, 159(1):142–151, 2011.

[57] Sture Forsén and Sara Linse. Cooperativity: over the hill. *Trends in Biochemical Sciences*, 20(12):495 – 497, 1995.

[58] Daniel L. Jones, Robert C. Brewster, and Rob Phillips. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216):1533–1536, 2014.

[59] Avigdor Eldar and Michael Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173, 2010.

[60] Ulrich Gerland and Terence Hwa. On the selection and evolution of regulatory DNA motifs. *Journal of Molecular Evolution*, 55(4):386–400, 2002.

[61] Johannes Berg, Stana Willmann, and Michael Lässig. Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology*, 4(1):42, 2004.

[62] Konstantin B. Zeldovich and Eugene I. Shakhnovich. Understanding protein evolution: from protein physics to Darwinian selection. *Annual Review of Physical Chemistry*, 59(1):105–127, 2008.

[63] Shyam K. Sharan, Lynn C. Thomason, Sergey G. Kuznetsov, and Donald L. Court. Recombineering: a homologous recombination-based method of genetic engineering. *Nature Protocols*, 4(2):206–223, 2009.

[64] Howard M. Salis, Ethan A. Mirsky, and Christopher A. Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, 27(10):946–950, 2009.

[65] Lynn C. Thomason, Nina Costantino, and Donald L. Court. *E. coli* genome manipulation by P1 transduction. *Current Protocols in Molecular Biology*, Chapter 1:Unit 1.17–1.17.8, 2007.

[66] Alfred Fernández-Castané, Claire E. Vine, Glòria Caminal, and Josep López-Santín. Evidencing the role of lactose permease in IPTG uptake by *Escherichia coli* in fed-batch high cell density cultures. *Journal of Biotechnology*, 157(3):391–398, 2012.

[67] Mitchell Lewis, Geoffrey Chang, Nancy C. Horton, Michele A. Kercher, Helen C. Pace, Maria A. Schumacher, Richard G. Brennan, and Ponzy Lu. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, 271(5253):1247–54, 1996.

[68] Holden T. Maecker, Aline Rinfret, Patricia D'Souza, Janice Darden, Eva Roig, Claire Landry, Peter Hayes, Josephine Birungi, Omu Anzala, Miguel Garcia, Alexandre Harari, Ian Frank, Ruth Baydo, Megan Baker, Jennifer Holbrook, Janet Ottinger, Laurie Lamoreaux, C. Lorrie Epling, Elizabeth Sinclair, Maria A. Suni, Kara Punt, Sandra Calarota, Sophia El-Bahi, Gailet Alter, Hazel Maila, Ellen Kuta, Josephine Cox, Clive Gray, Marcus Altfeld, Nolwenn Nougarede, Jean Boyer, Lynda Tussey, Timothy Tobery, Barry Bredt, Mario Roederer, Richard Koup, Vernon C. Maino, Kent Weinhold, Giuseppe Pantaleo, Jill Gilmour, Helen Horton, and Rafick P. Sekaly. Standardization of cytokine flow cytometry assays. *BMC Immunology*, 6(1):13, 2005.

[69] Kenneth Lo, Ryan Remy Brinkman, and Raphael Gottardo. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A*, 73A(4):321–332, 2008.

[70] Nima Aghaeepour, Greg Finak, The FlowCAP Consortium, The DREAM Consortium, Holger Hoos, Tim R. Mosmann, Ryan Brinkman, Raphael Gottardo, and Richard H. Scheuermann. Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods*, 10(3):228–238, 2013.

[71] Alexandra K. Gardino, Brian F. Volkman, Ho S. Cho, Seok-Yong Lee, David E. Wemmer, and Dorothee Kern. The NMR solution structure of BeF3-activated Spo0F reveals the conformational switch in a phosphorelay system. *Journal of Molecular Biology*, 331(1):245–254, 2003.

[72] Stephen Boulton and Giuseppe Melacini. Advances in NMR methods to map allosteric sites: From models to translation. *Chemical Reviews*, 116(11):6267–6304, 2016.

[73] Stefan Oehler, Siegfried Alberti, and Benno Müller-Hill. Induction of the *lac* promoter in the absence of DNA loops and the stoichiometry of induction. *Nucleic Acids Research*, 34(2):606–612, 2006.

[74] Mattias Rydenfelt, Hernan G. Garcia, Robert Sidney Cox, and Rob Phillips. The influence of promoter architectures and regulatory motifs on gene expression in *Escherichia coli*. *PLoS ONE*, 9(12):1–31, 2014.

[75] Alexander Schmidt, Karl Kochanowski, Silke Vedelaar, Erik Ahrné, Benjamin Volkmer, Luciano Callipo, Kèvin Knoops, Manuel Bauer, Ruedi Aebersold, and Matthias Heinemann. The quantitative and condition-dependent *Escherichia coli* proteome. *Nature Biotechnology*, 34(1):104–110, 2015.

[76] Chroma Technology Corporation. Chroma spectra viewer, 2016.

[77] Arthur D. Edelstein, Mark A. Tsuchida, Nenad Amodaj, Henry Pinkard, Ronald D. Vale, and Nico Stuurman. Advanced methods of microscope control using μManager software. *Journal of Biological Methods*, 1(2):10–10, 2014.

[78] David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society B: Biological Sciences*, 207(1167):187–217, 1980.

[79] Steven Frank. Input-output relations in biological systems: measurement, information and the Hill equation. *Biology Direct*, 8(1):31, 2013.

[80] James N. Weiss. The Hill equation revisited: uses and misuses. *The FASEB Journal*, 11(11):835–41, 1997.

Chapter 3

A SYSTEMATIC APPROACH FOR DISSECTING THE MOLECULAR MECHANISMS OF TRANSCRIPTIONAL REGULATION IN BACTERIA

A version of this chapter is in press as Nathan M. Belliveau, Stephanie L. Barnes, William T. Ireland, Daniel L. Jones, Michael Sweredoski, Annie Moradian, Sonja Hess, Justin B. Kinney, and Rob Phillips. A systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proceedings of the National Academy of Sciences*, In press, 2018.

Author contribution note: for this chapter, I (SB) assisted with experimental design, Sort-Seq sample processing, and writing the manuscript.

3.1 Introduction

The sequencing revolution has left in its wake an enormous challenge: the rapidly expanding catalog of sequenced genomes is far outpacing a sequence-level understanding of how the genes in these genomes are regulated. This ignorance extends from viruses to bacteria to archaea to eukaryotes. Even in *E. coli*, the model organism in which transcriptional regulation is best understood, we still have no indication if or how more than half of the genes are regulated (See Supplemental Figure 3.8; see also RegulonDB [1] or EcoCyc [2]). In other model bacteria such as *Bacillus subtilis, Caulobacter crescentus, Vibrio harveyii*, or *Pseudomonas aeruginosa*, far fewer genes have established regulatory mechanisms [3–5].

New approaches are needed for studying regulatory architecture in these and other bacteria. Chromatin immunoprecipitation and other high-throughput techniques are increasingly being used to study gene regulation in *E. coli* [6–11], but these methods are incapable of revealing either the nucleotide-resolution location of all functional transcription factor binding sites, or the way in which interactions between DNA-bound transcription factors and RNA polymerase modulate transcription. Although an arsenal of now classic genetic and biochemical methods have been developed for dissecting promoter function at individual bacterial promoters (reviewed in Minchin *et al.* [12]), these methods are not readily parallelized and often require purification of promoter-specific regulatory proteins.

In recent years a variety of massively parallel reporter assays have been developed for dissecting the functional architecture of transcriptional regulatory sequences in bacteria, yeast, and metazoans. These technologies have been used to infer biophysical models of well-studied loci, to characterize synthetic promoters constructed from known binding sites, and to search for new transcriptional regulatory sequences [13–19]. CRISPR assays have also shown promise for identifying longer range enhancer-promoter interactions in mammalian cells [20]. However, no approach for using massively parallel reporter technologies to decipher the functional mechanisms of previously uncharacterized regulatory sequences has yet been established.

Here we take a first step toward quantitative, multi-promoter dissection and describe a systematic approach for identifying the functional architecture of previously uncharacterized bacterial promoters at nucleotide resolution using a combination of genetic, functional, and biochemical measurements. First, a massively parallel reporter assay (Sort-Seq [13]) is performed on a promoter in multiple growth conditions in order to identify functional transcription factor binding sites. DNA affinity chromatography and mass spectrometry [21, 22] are then used to identify the regulatory proteins that recognize these sites. In this way one is able to identify both the functional transcription factor binding sites and cognate transcription factors in previously unstudied promoters. Subsequent massively parallel assays are then performed in gene-deletion strains to provide additional validation of the identified regulators. The reporter data thus generated is also used to infer sequence-dependent quantitative models of transcriptional regulation. In what follows, we first illustrate the overarching logic of our approach through application to four previously annotated promoters: *lacZYA*, *relBE*, *marRAB*, and *yebG*. We then apply this strategy to the previously uncharacterized promoters of *purT*, *xylE*, and *dgoRKADT*, demonstrating the ability to go from regulatory ignorance to explicit quantitative models of a promoter's input-output behavior.

3.2 Results

To dissect how a promoter is regulated, we begin by performing Sort-Seq [13]. As shown in Figure 3.1A, Sort-Seq works by first generating a library of cells, each of which contains a mutated promoter that drives expression of GFP from a low copy plasmid (5-10 copies per cell [23]) and provides a read-out of transcriptional state. We use fluorescence-activated cell sorting (FACS) to sort cells into multiple bins gated by their fluorescence level and then sequence the mutated plasmids from each bin. We found it sufficient to sort the libraries into four bins and generated data sets of about 0.5-2 million sequences across the sorted bins (Section 3.6, Figure 3.6A-D). To identify putative binding sites, we calculate 'expression shift' plots that show the average change in fluorescence when each position of the regulatory DNA is mutated (Figure 3.1B, top plot). Mutations to the DNA will in general disrupt binding of transcription factors [24], so regions with a positive shift are suggestive of binding by a repressor, while a negative shift suggests binding by an activator or RNA polymerase (RNAP).

The identified binding sites are further interrogated by performing information-based modeling with the Sort-Seq data. Here we generate energy matrix models [13, 25] that describe the sequence-dependent energy of interaction of a transcription factor at each putative binding site. For each matrix, we use a convention that the wild-type sequence is set to have an energy of zero (see example energy matrix in Figure 3.1B). Mutations that enhance binding are identified in blue, while mutations that weaken binding are identified in red. We also use these energy matrices to generate sequence logos [26] which provides a useful visualization of the sequence-specificity (see above matrix in Figure 3.1B).

In order to identify the putative transcription factors, we next perform DNA affinity chromatography experiments using DNA oligonucleotides containing the binding sites identified by Sort-Seq. Here we apply a stable isotopic labeling of cell culture (SILAC [27–30]) approach, which enables us to perform a second reference affinity chromatography that is simultaneously analyzed by mass spectrometry. We perform chromatography using magnetic beads with tethered oligonucleotides containing the putative binding site (Figure 3.1C). Our reference purification is performed identically, except that the binding site has been mutated away. The abundance of each protein is determined by mass spectrometry and used to calculate protein enrichment ratios, with the target transcription factor expected to exhibit a ratio greater than one. The reference purification ensures that non-specifically bound

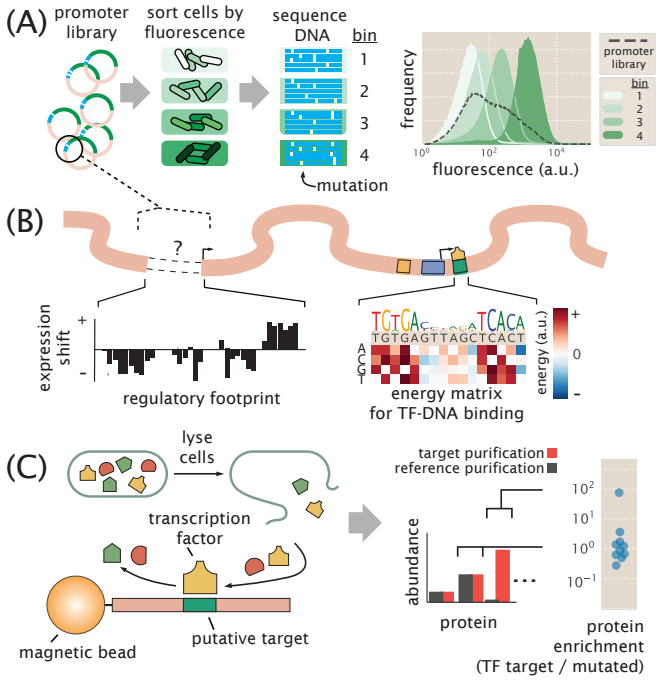

Figure 3.1: **Overview of approach to characterize transcriptional regulatory DNA, using Sort-Seq and mass spectrometry.** (A) Schematic of Sort-Seq. A promoter plasmid library is placed upstream of GFP and is transformed into cells. The cells are sorted into four bins by FACS and after regrowth, plasmids are purified and sequenced. The entire intergenic region associated with a promoter is included on the plasmid and a separate downstream ribosomal binding site sequence is used for translation of the *GFP* gene. The fluorescence histograms show the fluorescence from a library of the *rel* promoter and the resulting sorted bins. (B) Regulatory binding sites are identified by calculating the average expression shift due to mutation at each position. In the schematic, positive expression shifts are suggestive of binding by repressors, while negative shifts would suggest binding by an activator or RNAP. Quantitative models can be inferred to describe and further interrogate the associated DNA-protein interactions. An example energy matrix that describes the binding energy between an as yet unknown transcription factor to the DNA is shown. By convention, the wild-type nucleotides have zero energy, with blue squares identifying mutations that enhance binding (negative energy), and where red squares reduce binding (positive energy). The wild-type sequence is written above the matrix. (C) DNA affinity chromatography and mass spectrometry is used to identify the putative transcription factor (TF) for an identified repressor site. DNA oligonucleotides containing the target binding site are tethered to magnetic beads and used to purify the target transcription factor from cell lysate. Protein abundance is determined by mass spectrometry and a protein enrichment is calculated as the ratio in abundance relative to a second reference experiment where the target sequence is mutated away.

proteins will have a protein enrichment near one. This mass spectrometry data and the energy matrix models provide insight into the identity of each regulatory factor and potential regulatory mechanisms. In certain instances these insights then allow us to probe the Sort-Seq data further through additional information-based modeling using thermodynamic models of gene regulation. As further validation of binding by an identified regulator, we also perform Sort-Seq experiments in gene deletion strains, which should no longer show the associated positive or negative shift in expression at their binding site.

Sort-Seq recovers the regulatory features of well-characterized promoters

To first demonstrate Sort-Seq as a tool to discover regulatory binding sites *de novo* we began by looking at the promoters of *lacZYA* (*lac*), *relBE* (*rel*), and *marRAB* (*mar*). These promoters have been studied extensively [31–33] and provide a useful testbed of distinct regulatory motifs. To proceed we constructed libraries for each promoter by mutating their known regulatory binding sites. We begin by considering the *lac* promoter, which contains three *lac* repressor (LacI) binding sites, two of which we consider here, and a cyclic AMP receptor (CRP) binding site. It exhibits the classic catabolic switch-like behavior that results in diauxie when *E. coli* is grown in the presence of glucose and lactose sugars [31]. Here we performed Sort-Seq with cells grown in M9 minimal media with 0.5% glucose. The expression shifts at each nucleotide position are shown in Figure 3.2A, with annotated binding sites noted above the plot. The expression shifts reflect the expected regulatory role of each binding site, showing positive shifts for LacI and negative shifts for CRP and RNAP. The difference in magnitude at the two LacI binding sites likely reflect the different binding energies between these two binding site sequences, with LacI O3 having an *in vivo* dissociation constant that is almost three orders of magnitude weaker than the LacI O1 binding site [31, 34].

Next we consider the *rel* promoter that transcribes the toxin-antitoxin pair RelE and RelB. It is one of about 36 toxin-antitoxin systems found on the chromosome, with important roles in cellular physiology including cellular persistence [35]. When the toxin, RelE, is in excess of its cognate binding partner, the antitoxin RelB, the toxin causes cellular paralysis through cleavage of mRNA [36]. Interestingly, the antitoxin protein also contains a DNA binding domain and is a repressor of its own promoter [37]. We similarly performed Sort-Seq, with cells grown in M9 minimal media. The expression shifts are shown in Figure 3.2B and were consistent with binding by RNAP and RelBE. In particular, a positive shift was observed at the

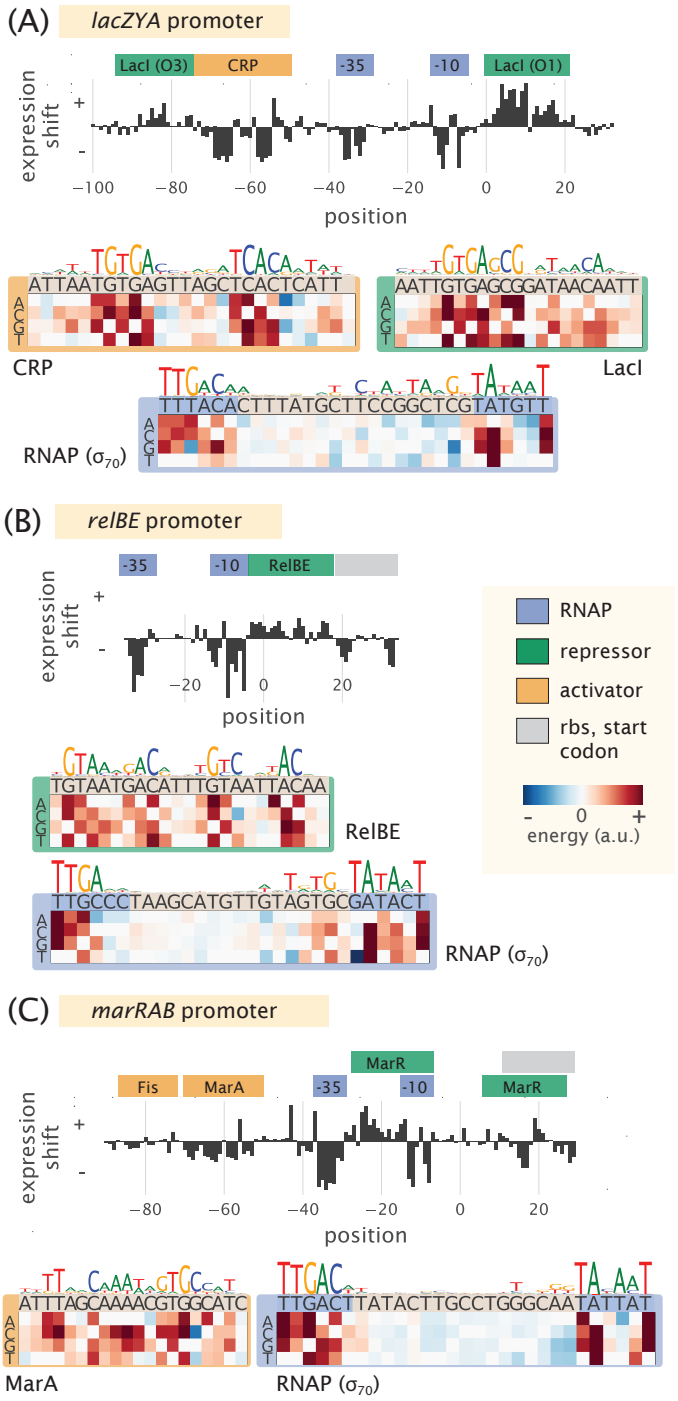

Figure 3.2: **Characterization of the regulatory landscape of the *lac*, *rel*, and *mar* promoters**. (A) Sort-Seq of the *lac* promoter. Cells were grown in M9 minimal media with 0.5% glucose at 37°C. Expression shifts and energy matrices are shown, with annotated binding sites for CRP (activator), RNAP (-10 and -35 subsites), and LacI (repressor) noted. (B) Sort-Seq of the *rel* promoter. Cells were also grown in M9 minimal media with 0.5% glucose at 37°C. The expression shifts identify the binding sites of RNAP and RelBE (repressor), and energy matrices and sequence logos are shown for these. (C) Sort-Seq of the *mar* promoter. Cells were grown in lysogeny broth (LB) at 30°C. The expression shifts identify the known binding sites of Fis and MarA (activators), RNAP, and MarR (repressor). Energy matrices and sequence logos are shown for MarA and RNAP.

binding site for RelBE, and the RNAP binding site mainly showed a negative shift in expression.

The third promoter, *mar*, is associated with multiple antibiotic resistance since its operon codes for the transcription factor MarA, which activates a variety of genes including the major multi-drug resistance efflux pump, ArcAB-tolC, and increases antibiotic tolerance [33]. The *mar* promoter is itself activated by MarA, SoxS, and Rob (via the so-called marbox binding site), and further enhanced by Fis, which binds upstream of this marbox [38]. Under standard laboratory growth it is under repression by MarR [33]. We found that the promoter's fluorescence was quite dim in M9 minimal media and instead grew libraries in lysogeny broth (LB) at 30°C [39]. Again, the different features in the expression shift plot (Figure 3.2C) appeared to be consistent with the noted binding sites. One exception was that the downstream MarR binding site was not especially apparent. Both positive and negative expression shifts were observed along its binding site, which may be due to overlap with other features present including the native ribosomal binding site. There have also been reported binding sites for CRP, Cra, CpxR/CpxA, and AcrR [1]. However the studies associated with these annotations either required over-expression of the associated transcription factor, were computationally predicted, or demonstrated through *in vitro* assays and not necessarily expected under the growth condition considered here.

While each promoter qualitatively showed the expected regulatory behavior in each expression shift plot, it was important to show that we could also recover the quantitative features of binding by each transcription factor. Here we inferred energy matrices and associated sequence logos for the binding sites of RNAP, LacI, CRP, RelBE, MarA, and Fis. These are shown in Figure 3.2A-C and in Supplemental Figure 3.11, and indeed, the matrices agreed well with those generated from known genomic binding sites for each transcription factor (Pearson correlation coefficient r=0.5-0.9; see Supplemental Section 3.7).

For the repressors RelBE and MarR, there was no data available that characterized their sequence specificity with which to compare against. Here, instead, we validated our data by performing Sort-Seq in strains where the *relBE* or *marR* genes were deleted. In each case this resulted in a loss of the expression shift associated with binding by these repressors (Figure 3.3) and an inability of the energy matrices to explain the data in the deletion strain (Supplemental Section 3.11, Figure 3.14), suggesting that the observed features in the wild-type strain data are due to binding

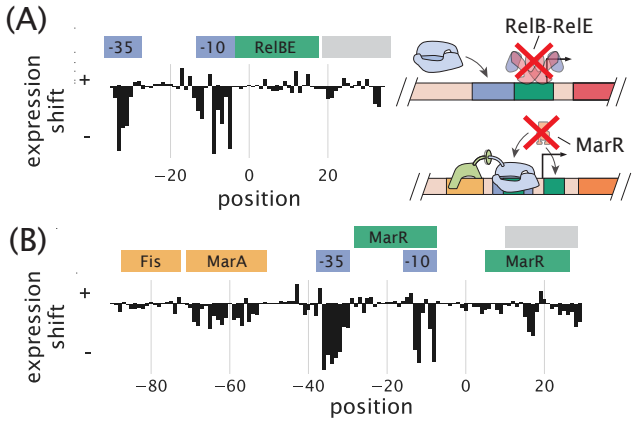

Figure 3.3: **Expression shifts reflect binding by regulatory proteins.** (A) Expression shifts for the *rel* promoter, but in a Δ*rel* genetic background. Cells were grown in conditions identical to Figure 3.2B but do not show a positive expression shift across the entire RelBE binding site. (B) Expression shifts for the *mar* promoter, but in a Δ*marR* genetic background. The positive expression shift observed where MarR is expected to bind is no longer observed. Binding site annotations are identified in blue for RNAP sites, green for repressor sites, yellow for activator sites, and gray for ribosomal binding site and start codons. These annotations refer to the binding sites noted on RegulonDB that were observed in the Sort-Seq data.

by these transcription factors.

Identification of transcription factors with DNA affinity chromatography and quantitative mass spectrometry

It was next important to show that DNA affinity chromatography could be used to identify transcription factors in *E. coli*. In particular, a challenge arises in identifying transcription factors in most organisms due to their very low abundance. In *E. coli* the cumulative distribution in protein copy number shows that more than half have a copy number less than 100 per cell, with 90% having copy number less than 1,000 per cell. This is several orders of magnitude below that of many other cellular proteins [40].

We began by applying the approach to known binding sites for LacI and RelBE. For LacI, which is present in *E. coli* in about 10 copies per cell, we used the strongest binding site sequence, Oid (*in vivo* $K_d \approx 0.05\,nM$), and the weakest natural operator sequence, O3 (*in vivo* $K_d \approx 110\,nM$) [31, 34, 41]. In Figure 3.4A we plot the protein enrichments from each transcription factor identified by mass spectrometry. LacI was found with both DNA targets, with fold enrichment greater than 10 in each case, and significantly higher than most of the proteins detected (indicated by the

shaded region, which represents the 95% probability density region of all proteins detected, including non-DNA binding proteins). Purification of LacI with about 10 copies per cell using the weak O3 binding site sequence are near the limit of what would be necessary for most *E. coli* promoters.

To ensure this success was not specific to LacI, we also applied chromatography to the RelBE binding site. RelBE provides an interesting case since the strength of binding by RelB to DNA is dependent on whether RelE is bound in complex to RelB (with at least a 100 fold weaker dissociation constant reported in the absence of RelE [42, 43]). As shown in Figure 3.4B, we found over 100 fold enrichment of both proteins by mass spectrometry. To provide some additional intuition into these results we also considered the predictions from a statistical mechanical model of DNA binding affinity (See Supplemental Section 3.8). As a consequence of performing a second reference purification, we find that fold enrichment should mostly reflect the difference in binding energy between the DNA sequences used in the two purifications, and be much less dependent on whether the protein was in low or high abundance within the cell. This appeared to be the case when considering other *E. coli* strains with LacI copy numbers between about 10 and 1,000 copies per cell (Supplemental Figure 3.12). Further characterization of the measurement sensitivity and dynamic range of this approach is noted in Supplemental Section 3.9.

Sort-Seq discovers regulatory architectures in unannotated regulatory regions
Given that more than half of the promoters in *E. coli* have no annotated transcription factor binding sites in RegulonDB, we narrowed our focus by using several high-throughput studies to identify candidate genes to apply our approach [44, 45]. The work by Schmidt *et al.* [45] in particular measured the protein copy number of about half the *E. coli* genes across 22 distinct growth conditions. Using this data, we identified genes that had substantial differential gene expression patterns across growth conditions, thus hinting at the presence of regulation and even how that regulation is elicited by environmental conditions (see further details in Supplemental Section 3.5 and Supplemental Figure 3.9A-C).

On the basis of this survey, we chose to investigate the promoters of *purT*, *xylE*, and *dgoRKADT*. To apply Sort-Seq in a more exploratory manner, we considered three 60 bp mutagenized windows spanning the intergenic region of each gene. While it is certainly possible that regulatory features will lie outside of this window, a search

of known regulatory binding sites suggests that this should be sufficient to capture just over 70% of regulatory features in *E. coli* and provide a useful starting point (Supplemental Figure 3.13).

The *purT* promoter contains a simple repression architecture and is repressed by PurR

The first of our candidate promoters is associated with expression of *purT*, one of two genes found in *E. coli* that catalyze the third step in *de novo* purine biosynthesis [46, 47]. Due to a relatively short intergenic region, about 120 bp in length that is shared with a neighboring gene *yebG*, we also performed Sort-Seq on the *yebG* promoter (oriented in the opposite direction [48]; see schematic in Figure 3.5A). To begin our exploration of the *purT* and *yebG* promoters, we performed Sort-Seq with cells grown in M9 minimal media with 0.5% glucose. The associated expression shift plots are shown in Figure 3.5A. While we performed Sort-Seq on a larger

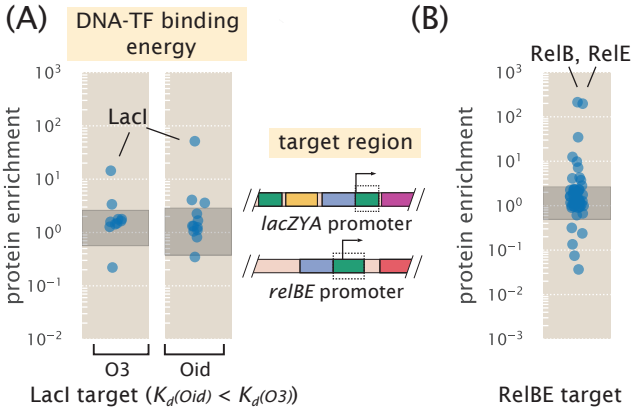

Figure 3.4: **DNA affinity purification and identification of LacI and RelBE by mass spectrometry using known target binding sites.** (A) Protein enrichment using the weak O3 binding site and strong synthetic Oid binding sites of LacI. LacI was the most significantly enriched protein in each purification. The target DNA region was based on the boxed area of the *lac* promoter schematic, but with the native O1 sequence replaced with either O3 or Oid. Data points represent average protein enrichment for each detected transcription factor, measured from a single purification experiment. (B) For purification using the RelBE binding site target, both RelB and its cognate binding partner RelE were significantly enriched. Data points show the average protein enrichment from two purification experiments. The target binding site is similarly shown by the boxed region of the *rel* promoter schematic. Data points in each purification show the protein enrichment for detected transcription factors. The gray shaded regions show where 95% of all detected protein ratios were found.

region than shown for each promoter, we only plot the regions where regulation was apparent.

For the *yebG* promoter, the features were largely consistent with prior work, containing a binding sites for LexA and RNAP. However, we found that the RNAP binding site is shifted 9 bp downstream from what was identified previously. The previously annotated binding site was based on a computational search [48] and not confirmed experimentally. We were also able to confirm that the *yebG* promoter was induced in response to DNA damage by repeating Sort-Seq in the presence of mitomycin C (a potent DNA cross-linker known to elicit the SOS response and proteolysis of LexA [49]; see Supplemental Figure 3.15A, B, and D).

Given the role of *purT* in the synthesis of purines, and the tight control over purine concentrations within the cell [46], we performed Sort-Seq of the *purT* promoter in the presence or absence of the purine, adenine, in the growth media. In growth without adenine (Figure 3.5A, right plot), we observed two negative regions in the expression shift plot. Through inference of an energy matrix, these two features were identified as the -10 and -35 regions of an RNAP binding site. While these two features were still present upon addition of adenine, as shown in Figure 3.5B, this growth condition also revealed a putative repressor site between the -35 and -10 RNAP binding sites, indicated by a positive shift in expression (green annotation).

Following our strategy to find not only the regulatory sequences, but also their associated transcription factors, we next applied DNA affinity chromatography using this putative binding site sequence. In our initial attempt however, we were unable to identify any substantially enriched transcription factor (Supplemental Figure 3.15C). With repression observed only when cells were grown in the presence of adenine, we reasoned that the transcription factor may require a related ligand in order to bind the DNA, possibly through an allosteric mechanism. Importantly, we were able to infer an energy matrix to the putative repressor site whose sequence-specificity matched that of the well-characterized repressor, PurR (r=0.82; see Supplemental Figure 3.11). We also noted ChIP-chip data of PurR that suggests it might bind within this intergenic region [47]. We therefore repeated the purification in the presence of hypoxanthine, which is a purine derivative that also binds PurR [50]. As shown in Figure 3.5C, we now observed a substantial enrichment of PurR with this putative binding site sequence. As further validation, we performed Sort-Seq once more in the adenine-rich growth condition, but in a $\Delta purR$ strain. In the absence of PurR, the putative repressor binding site disappeared (Figure 3.5D), which is consistent

with PurR binding at this location.

In Figure 3.5E we summarize the regulatory features between the coding genes of *purT* and *yebG*, including the new features identified by Sort-Seq. With the appearance of a simple repression architecture [51] for the *purT* promoter, we extended our analysis by developing a thermodynamic model to describe repression by PurR. This enabled us to infer the binding energies of RNAP and PurR in absolute k_BT energies [52], and we show the resulting model in Figure 3.5E (see additional details in Supplemental Section 3.12).

The *xylE* operon is induced in the presence of xylose, mediated through binding of XylR and CRP

The next unannotated promoter we considered was associated with expression of *xylE*, a xylose/proton symporter involved in uptake of xylose. From our analysis of the Schmidt *et al.* [45] data, we found that *xylE* was sensitive to xylose and proceeded by performing Sort-Seq in cells grown in this carbon source. Interestingly, the promoter exhibited essentially no expression in other media (see Schmidt *et al.* [45], and Supplemental Figure 3.15E). We were able to locate the RNAP binding site between -80 bp and -40 bp relative to the *xylE* gene (Figure 3.6A, annotated in blue). In addition, the entire region upstream of the RNAP appeared to be involved in activating gene expression (annotated in orange in Figure 3.6A), suggesting the possibility of multiple transcription factor binding sites.

We applied DNA affinity chromatography using a DNA target containing this entire upstream region. Due to the stringent requirement for xylose to be present for any measurable expression, xylose was supplemented in the lysate during binding with the target DNA. In Figure 3.6B we plot the enrichment ratios from this purification and find XylR to be most significantly enriched. From an energy matrix inferred for the entire region upstream of the RNAP site, we were able to identify two correlated 15 bp regions (dark yellow shaded regions in Figure 3.6C; Pearson correlation r = 0.74 between energy matrices from each binding site). Mutations of the XylR protein have been found to diminish transport of xylose [53], which in light of our result, may be due in part to a loss of activation and expression of this xylose/proton symporter. This is in addition to the loss of activation expected by XylR of the high-affinity xylose uptake system XylFHG [53]. These binding sites were also similar to those found on two other promoters known to be regulated by XylR (*xylA* and *xylF* promoters), whose promoters also exhibit tandem XylR binding sites and strong

binding energy predictions with our energy matrix (Supplemental Figure 3.15F).

Within the upstream activator region in Figure 3.6A there still appeared to be a binding site unaccounted for with these tandem XylR binding sites. From the energy matrix, we were further able to identify a binding site for CRP, which is noted upstream of the XylR binding sites in Figure 3.6C. While we did not observe a significant enrichment of CRP in our protein purification, the most energetically favorable sequence predicted by our model, TGCGACCNAGATCACA, closely matches the CRP consensus sequence of TGTGANNNNNNTCACA. In contrast to the *lac* promoter, binding by CRP here appears to depend more on the right half of the binding site sequence. CRP is known to activate promoters by multiple mechanisms [54], and CRP binding sites have been found adjacent to the activators XylR and AraC [53, 55], in line with our result. While further work will be needed to characterize the specific regulatory mechanism here, it appears that activation of RNAP is mediated by both CRP and XylR and we summarize this result in Figure 3.6D (and consider it further in Supplemental Section 3.12).

The *dgoRKADT* promoter is auto-repressed by DgoR, with transcription mediated by class II activation by CRP

As a final illustration of the approach developed here, we considered the unannotated promoter of *dgoRKADT*. The operon codes for D-galactonate-catabolizing enzymes; D-galactonate is a sugar acid that has been found as a product of galactose metabolism [56]. We began by measuring expression from a non-mutagenized *dgoRKADT* promoter reporter to glucose, galactose, and D-galactonate. Cells grown in galactose exhibited higher expression than in glucose, as found by Schmidt *et al.* [45], and even higher expression when cells were grown in D-galactonate (Supplemental Figure 3.16A). This likely reflects the physiological role provided by the genes of this promoter, which appear necessary for metabolism of D-galactonate. We therefore proceeded by performing Sort-Seq with cells grown in either glucose or D-galactonate, since these appeared to represent distinct regulatory states, with expression low in glucose and high in D-galactonate. Expression shift plots from each growth conditions are shown in Figure 3.7A.

We begin by considering the results from growth in glucose (Figure 3.7A, top plot). Here we identified an RNAP binding site between -30 bp and -70 bp, relative to the native start codon for *dgoR* (Supplemental Figure 3.16B). Another distinct feature was a positive expression shift in the region between -140 bp and -110

bp, suggesting the presence of a repressor binding site. Applying DNA affinity chromatography using this target region, we observed an enrichment of DgoR (Figure 3.7B), suggesting that the promoter is indeed under repression, and regulated by the first coding gene of its transcript. As further validation of binding by DgoR, the positive shift in expression was no longer observed when Sort-Seq was repeated in a $\Delta dgoR$ strain (Figure 3.7D and Supplemental Figure 3.16C). We also were able to identify additional RNAP binding sites that were not apparent due to binding by DgoR. While only one RNAP -10 motif is clearly visible in the sequence logo shown Figure 3.7C (top sequence logo; TATAAT consensus sequence), we used simulations to demonstrate that the entire sequence logo shown can be explained by the convolution of three overlapping RNAP binding sites (See Supplemental Figure 3.16).

Next we consider the D-galactonate growth condition (Figure 3.7A, bottom plot). Like in the expression shift plot for the $\Delta dgoR$ strain grown in glucose, we no longer observe the positive expression shift between -140 bp and -110 bp. While there are still several positions between -120 bp and -100 bp that are still positive, this can be attributed to a non-optimal -10 binding site sequence for RNAP (wild-type TACATT, Figure 3.7C). The loss of the repressive feature therefore suggests that DgoR may be induced by D-galactonate or a related metabolite. However, in comparison with the expression shifts in the $\Delta dgoR$ strain grown in glucose, there were some notable differences in the region between -160 bp and -140 bp. Here we find evidence for another CRP binding site. The sequence logo identifies the sequence TGTGA (Figure 3.7C, bottom logo), which matches the left side of the CRP consensus sequence. In contrast to the *lac* and *xylE* promoters however, the right half of the binding site directly overlaps with where we would expect to find a -35 RNAP binding site. This type of interaction by CRP has been previously observed and is defined as class II CRP dependent activation [54], though this sequence-specificity has not been previously described.

In order to isolate and better identify this putative CRP binding site we repeated Sort-Seq in *E. coli* strain JK10, grown in 500 μM cAMP. Strain JK10 lacks adenlyate cyclase (*cyaA*) and phosphodiesterase (*cpdA*), which are needed for cAMP synthesis and degradation, respectively, and is thus unable to control intracellular cAMP levels necessary for activation by CRP (derivative of TK310 [41]). Growth in the presence of 500 μM cAMP provided strong induction from the *dgoRKADT* promoter and resulted in a sequence logo at the putative CRP binding site that even more clearly

resembled binding by CRP (Supplemental Figure 3.16E). This is likely because expression is now dominated by the CRP activated RNAP binding site. Importantly, this data allowed us to further infer the interaction energy between CRP and RNAP, which we estimate to be -7.3 k_BT (further detailed in Supplemental Section 3.12). We summarize the identified regulatory features in Figure 3.7E.

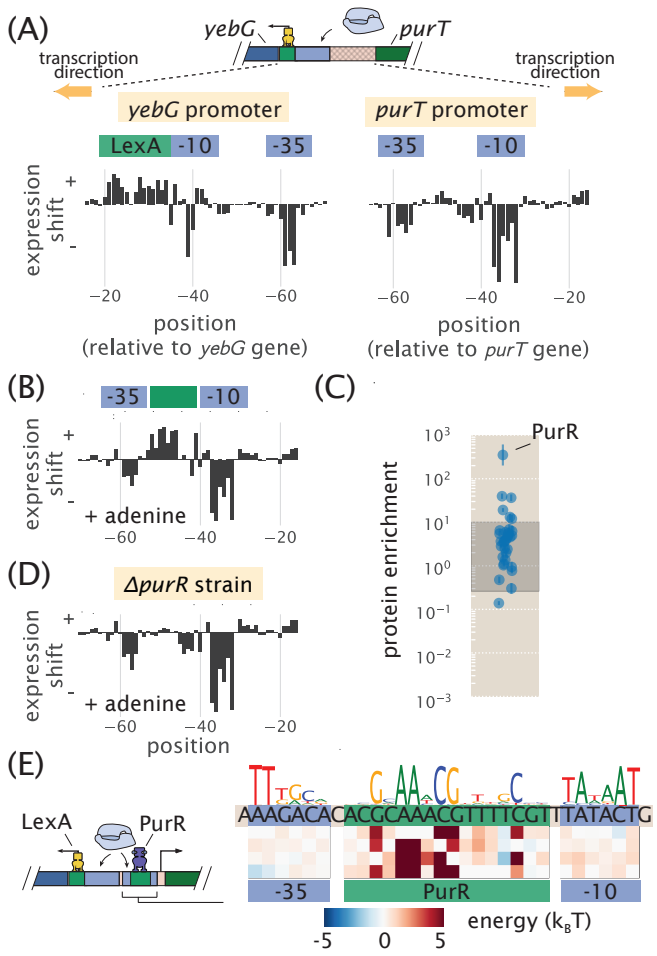

Figure 3.5: **Sort-Seq distinguishes directional regulatory features and uncovers the regulatory architecture of the *purT* promoter.** (A) A schematic is shown for the approximately 120 bp region between the *yebG* and *purT* genes, which code in opposite directions. Expression shifts are shown for 60 bp regions where regulation was observed for each promoter, with positions noted relative to the start codon of each native coding gene. Cells were grown in M9 minimal media with 0.5% glucose. The -10 and -35 RNAP binding sites of the *purT* promoter were determined through inference of an energy matrix and are identified in blue. (B) Expression shifts for the *purT* promoter, but in M9 minimal media with 0.5% glucose supplemented with adenine (100 μg/ml). A putative repressor site is annotated in green. (C) DNA affinity chromatography was performed using the identified repressor site and protein enrichment values for transcription factors are plotted. Cell lysate was produced from cells grown in M9 minimal media with 0.5 % glucose. Binding was performed in the presence of hypoxanthine (10 μg/ml). Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates, and the gray shaded region represents 95% probability density region of all protein detected. (D) Identical to (B) but performed with cells containing a Δ*purR* genetic background. (E) Summary of regulatory binding sites and transcription factors that bind within the intergenic region between the genes of *yebG* and *purT*. Energy weight matrices and sequence logos are shown for the PurR repressor and RNAP binding sites. Data was fit to a thermodynamic of simple repression, yielding energies in units of k_BT.

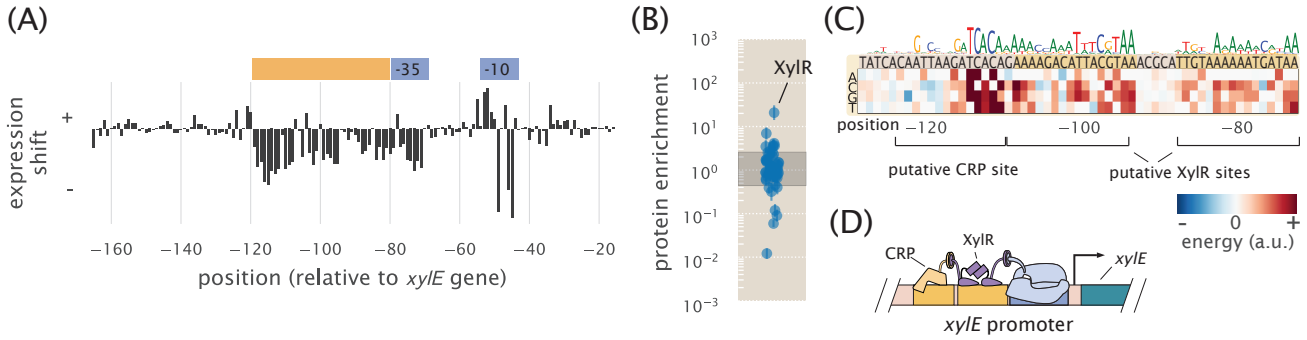

Figure 3.6: **Sort-Seq identifies a set of activator binding sites that drive expression of RNAP at the *xylE* promoter**. (A) Expression shifts are shown for the *xylE* promoter, with Sort-Seq performed on cells grown in M9 minimal media with 0.5% xylose. The -10 and -35 regions of an RNAP binding site (blue) and a putative activator region (orange) are annotated. (B) DNA affinity chromatography was performed using the putative activator region and protein enrichment values for transcription factors are plotted. Cell lysate was generated from cells grown in M9 minimal media with 0.5% xylose and binding was performed in the presence of xylose supplemented at the same concentration as during growth. Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates. The gray shaded region represents 95% probability density region of all proteins detected. (C) An energy matrix was inferred for the region upstream of the RNAP binding site. The associated sequence logo is shown above the matrix. Two binding sites for XylR were identified (see also Supplemental Section 3.7, Supplemental Figure 3.11 and Supplemental Figure 3.15F) along with a CRP binding site. (D) Summary of regulatory features identified at *xylE* promoter, with the identification of an RNAP binding site and tandem binding sites for XylR and CRP.

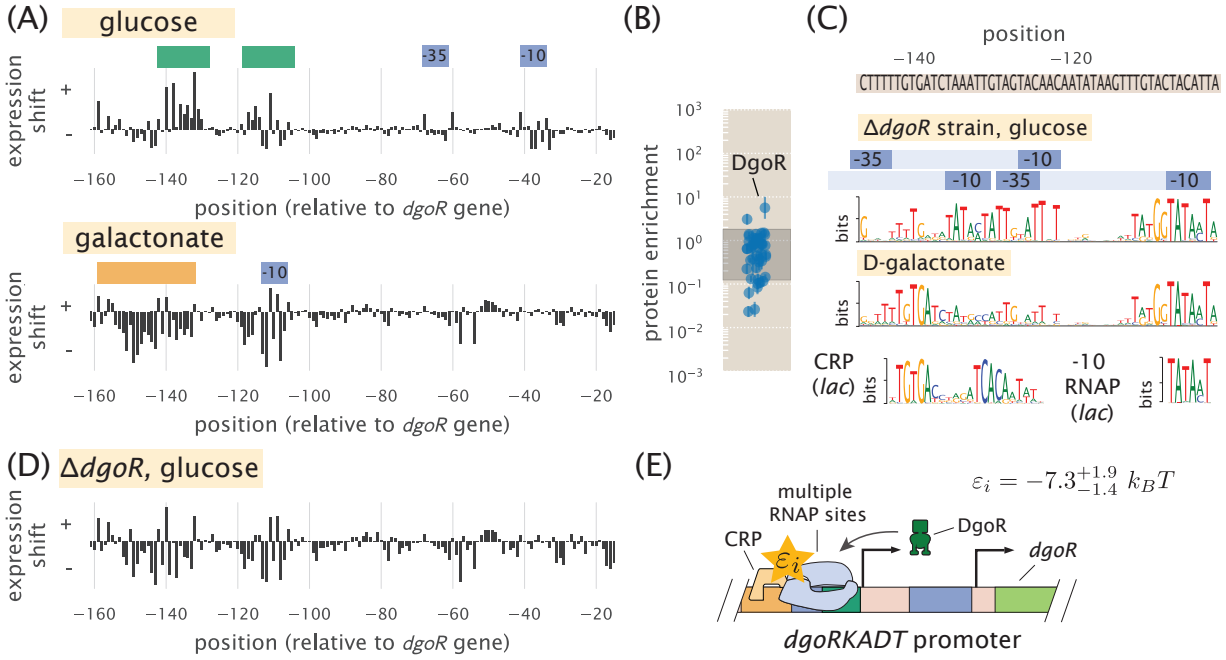

Figure 3.7: **The *dgoRKADT* promoter is induced in the presence of D-galactonate due to loss of repression by DgoR and activation by CRP.** (A) Expression shifts due to mutating the *dgoRKADT* promoter are shown for cells grown in M9 minimal media with either 0.5% glucose (top) or 0.23% D-galactonate (bottom). Regions identified as RNAP binding sites (-10 and -35) are shown in blue and putative activator and repressor binding sites are shown in orange and green, respectively. (B) DNA affinity purification was performed targeting the region between -145 to -110 of the *dgoRKADT* promoter. The transcription factor DgoR was found most enriched among the transcription factors plotted. Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates, and the gray shaded region represents 95% probability density region of all proteins detected. (C) Sequence logos were inferred for the most upstream 60 bp region associated with the upstream RNAP binding site annotated in (A). Multiple RNAP binding sites were identified using Sort-Seq data performed in a Δ*dgoR* strain, grown in M9 minimal media with 0.5% glucose. (further detailed in Supplemental Figure 3.16). Below this, a sequence logo was also inferred using data from Sort-Seq performed on wild-type cells, grown in D-galactonate, identifying a CRP binding site (class II activation [54]). (D) Expression shifts are shown for the *dgoRKADT* promoter when performed in a Δ*dgoR* genetic background, grown in 0.5% glucose. This resembles growth in D-galactonate, suggesting D-galactonate may act as an inducer for DgoR. (E) Summary of regulatory features identified at *dgoRKADT* promoter, with the identification of multiple RNAP binding sites, and binding sites for DgoR and CRP. The interaction energy between CRP and RNAP, ε_i, was inferred to be $-7.3^{+1.9}_{-1.4} k_BT$, where the superscripts and subscripts represent the upper and lower bounds of the 95$^{\text{th}}$ percentile of the parameter value distribution.

3.3 Discussion

We have established a systematic procedure for dissecting the functional mechanisms of previously uncharacterized regulatory sequences in bacteria. A massively parallel reporter assay, Sort-Seq [13], is used to first elucidate the locations of functional transcription factor binding sites. DNA oligonucleotides containing these binding sites are then used to enrich the cognate transcription factors and identify them by mass spectrometry analysis. Information-based modeling and inference of energy matrices that describe the DNA binding specificity of regulatory factors provide further quantitative insight into transcription factor identity and the growth condition dependent regulatory architectures.

To validate this approach we examined four previously annotated promoters of *lac*, *rel*, *mar*, and *yebG*, with our results consistent with established knowledge [13, 31, 33, 34, 39, 43]. Importantly, we find that DNA affinity chromatography experiments on these promoters were highly sensitive. In particular, LacI was unambiguously identified with the weak O3 binding site, even though LacI is present in only about 10 copies per cell [34]. Emboldened by this success, we then studied promoters having little or no prior regulatory annotation: *purT*, *xylE*, and *dgoR*. Here our analysis led to a collection of new regulatory hypotheses. For the *purT* promoter, we identified a simple repression architecture [51], with repression by PurR. The *xylE* promoter was found to undergo activation only when cells are grown in xylose, likely due to allosteric interaction between the activator XylR and xylose, and activation by CRP [53, 55]. Finally, in the case of *dgoR*, the base-pair resolution allowed us to tease apart overlapping regulatory binding sites, identify multiple RNAP binding sites along the length of the promoter, and infer further quantitative detail about the interaction between the newly identified binding sites for CRP and RNAP. We view these results as a critical first step in the quantitative dissection of transcriptional regulation, which will ultimately be needed for a predictive understanding of how such regulation works.

While our results show the successful identification of regulatory binding sites and regulatory mechanism at previously unannotated promoters, there also remain important challenges. The uncharacterized genes were selected based upon genome-wide studies [44, 45] and indeed, the hints of regulation in these data were a necessary part of our strategy to systematically dissect each promoter. Data sets that quantitate protein abundance across a number of growth conditions, like those available in *E. coli* [45] and yeast [57], or alternatively, transcript abundance using RNA-seq, will

provide an important starting point for the dissection of regulatory mechanism in other bacteria.

An important aspect of the presented approach is that it can be applied to any promoter sequence, and there are a number of ways that throughput can be increased further. Microarray-synthesized promoter libraries and measurement of expression from barcoded transcripts using RNA-seq, instead of flow cytometry, can be used to allow multiple loci to be studied simultaneously [14, 18]. Landing pad technologies for chromosomal integration [58–60] should enable massively parallel reporter assays to be performed in chromosomes instead of on plasmids. Techniques that combine these assays with transcription start site readout [61] may provide additional resolution, further allowing the molecular regulators of overlapping RNAP binding sites to be deconvolved, or the contributions from separate RNAP binding sites, like those observed on the *dgoR* promoter, to be better distinguished. As the number of regulatory regions under study increases, it will also be important to develop additional analysis tools that provide automated identification of regulatory binding sites.

In order to identify transcription factors across many target binding sites, DNA-affinity chromatography samples can be further mutliplexed using isobaric labeling strategies [62, 63]. Continued performance improvements in mass spectrometer sensitivity and sample processing [64–66] will also make this assay less onerous to apply across many targets and different binding conditions. This will be especially important for situations where the data suggests a small-molecule effector might be acting to modulate binding of the transcription factor to its target sequence, requiring multiple binding conditions to be tested. Performing reporter assays in transcription factor deletion strains will continue to play an important role in promoter dissection, as we have shown for a variety of the promoters, and provide a secondary means with which to identify and validate binding sites. Genome-wide knockout libraries are now available for a wide variety of bacteria [67–72], and it is now possible to perform genetic perturbations using CRISPRi [73, 74] that should open up the possibility of applying such perturbation strategies more easily in less-studied organisms.

Although our work was directed toward regulatory regions of *E. coli*, there are no intrinsic limitations that restrict the analysis to this organism. Rather, most bacteria contain small intergenic regions several hundred base-pairs in length that make this approach especially suitable. The sequence specificity of most characterized prokaryotic transcription factors [75, 76], and the sigma factors that allow RNAP to

recognize each promoter [54, 77], suggests that this approach will permit regulatory dissection in any bacterium that supports efficient transformation by plasmids. And although we have focused on bacteria, our general strategy should be feasible for dissecting regulation in a number of eukaryotic systems – including human cell culture – using massively parallel reporter assays [14–16] and DNA-mediated protein pull-down methods [21, 22] that have already been established.

3.4 Methods

Bacterial strains

All *E. coli* strains used in this work were derived from K-12 MG1655, with deletion strains generated by the lambda red recombinase method [78]. In the case of deletions for *lysA* ($\Delta lysA$::kan), *purR* ($\Delta purR$::kan), and *xylE* ($\Delta xylE$::kan), strains were obtained from the Coli Genetic Stock Center (CGSC, Yale University, CT, USA) and transferred into a fresh MG1655 strain using P1 transduction. The others were generated in house and include the following deletion strains: $\Delta lacIZYA$, $\Delta relBE$::kan, $\Delta marR$::kan, $\Delta dgoR$::kan. Details on strain construction are provided in Supplemental Section 3.13.

Sort-Seq

Mutagenized single-stranded oligonucleotide pools were purchased from Integrated DNA Technologies (Coralville, IA). Library oligonucleotides were PCR amplified and inserted into the PCR amplified plasmid backbone (i.e. vector) of pJK14 (SC101 origin) [13] by Gibson assembly and electroporated into cells following drop dialysis in water. Cell libraries were then grown to saturation in LB and then diluted 1:10,000 into the appropriate growth media for the promoter under consideration. A Beckman Coulter MoFlo XDP cell sorter was used to sort cells by fluorescence, with 500,000 cells collected into each of the four bins. Sorted cells were then re-grown overnight in 10 ml of LB media, under kanamycin selection. The plasmid in each bin were miniprepped (Qiagen, Germany) following overnight growth and PCR was used to amplify the mutated region from each plasmid for Illumina sequencing. See Supplemental Section 3.13 for additional details on library construction and Sort-Seq, and Section H on calculating expression shift plots and energy matrices.

DNA affinity chromatography and LC-MS/MS

SILAC labeling [27, 28, 30] was implemented by growing cells (MG1655 $\Delta lysA$) in either the stable isotopic form of lysine ($^{13}C_6H_{14}{}^{15}N_2O_2$) or natural form. See Supplemental Section 3.13 for details on lysate preparation.

DNA affinity chromatography was performed by incubating cell lysate (~150 mg/ml protein) with magnetic beads (Dynabeads MyOne T1, ThermoFisher, Waltham, MA) containing tethered DNA (streptavidin-biotin linkage). Single-stranded DNA was purchased from Integrated DNA Technologies with the biotin modification on the 5' end of the oligonucleotide sense strand. Cell lysates were incubated on a rotating wheel with the DNA tethered beads overnight at 4°C. Elution was

achieved by cleaving the DNA with the restriction enzyme PstI, and samples were then prepared for mass spectrometry by in-gel digestion with endoproteinase Lys-C. Liquid chromatography tandem-mass spectrometry (LC-MS/MS) experiments were carried out as previously described [79] and further detailed in Supplemental Section 3.13. Thermo RAW files were processed using MaxQuant (v. 1.5.3.30) [80].

Code availability and data analysis

All code used for processing data and plotting, as well as the final processed data, plasmid sequences, and primer sequences can be found on our GitHub repository (`https://www.github.com/RPGroup-PBoC/sortseq_belliveau`; DOI: `https://doi.org/10.5281/zenodo.1184169`). Thermo RAW files for mass spectrometry are available on the jPOSTrepo repository [81] under accession code PXD007892. Sort-Seq sequencing files are available on the Sequence Read Archive (accession code SRP121362; will be made available upon publication).

3.5 Supplemental Information: Identification of unannotated promoters in *E. coli* with growth-dependent differential expression.

Here we briefly describe how the unannotated promoters of the main text (*purT*, *xylE*, and *dgoR*) were chosen. Figure 3.8 summarizes the current state of regulatory knownledge in *E. coli* and those promoters considered in this work. Here, we parse the database RegulonDB that lists all known regulatory reatures in *E. coli*, with the striking finding that more than half the operons lack any annotated transcription factor binding sites (denoted by red lines). To identify candidate promoters with which to apply Sort-Seq, we made use of a variety of genome-wide datasets [40, 44, 45]. Specifically, in the case of the *purT* promoter, network inference approaches [44] led us to a number of unannotated genes that appeared to be sensitive to purine (others included *yieH* and *adeP*). Since the *purT* promoter lacked any experimental characterization and with ChIP-chip data suggesting PurR may be involved [47], it appeared to be a good starting point with which to apply our approach.

The promoters of *xylE* and *dgoR*, were identified from a recent study by Schmidt *et al.* [45]. They measured the copy number per cell of more than 2,300 proteins (about 55% of the *E. coli* proteome) across 22 growth conditions. These conditions included different carbon sources, temperature and pH, growth phase, media, and growth in chemostats. This provided us with a rich set of measurements with which to identify unannotated promoters where a particular growth condition influenced expression and may be under transcriptional regulation. The rest of this section describes how that data was used to identify candidate promoters.

In order to identify candidate genes using the mass spectrometry data, we ranked each protein based on its copy number in a particular growth condition, divided by the average copy number across the 22 conditions. Regulated proteins should be among those that exhibit a large change in copy number in one or a few growth conditions. As a confirmation of this, among the proteins with known regulation, we came across the GalE protein which was found to have significantly higher expression when cells were grown in galactose (Figure 3.9A). GalE is involved in galactose catabolism, and its expression is known to increase due to loss of repression of the *galE* promoter when cells were grown in galactose [82, 83]. Among promoters without any known regulation, we show the expression of DgoD in Figure 3.9B for several different carbon sources. Cells grown in galactose showed much higher expression of the DgoD gene, with about 675 copies per cell, compared to at most 15 copies per cell across the other growth conditions. This is only one of many examples where a

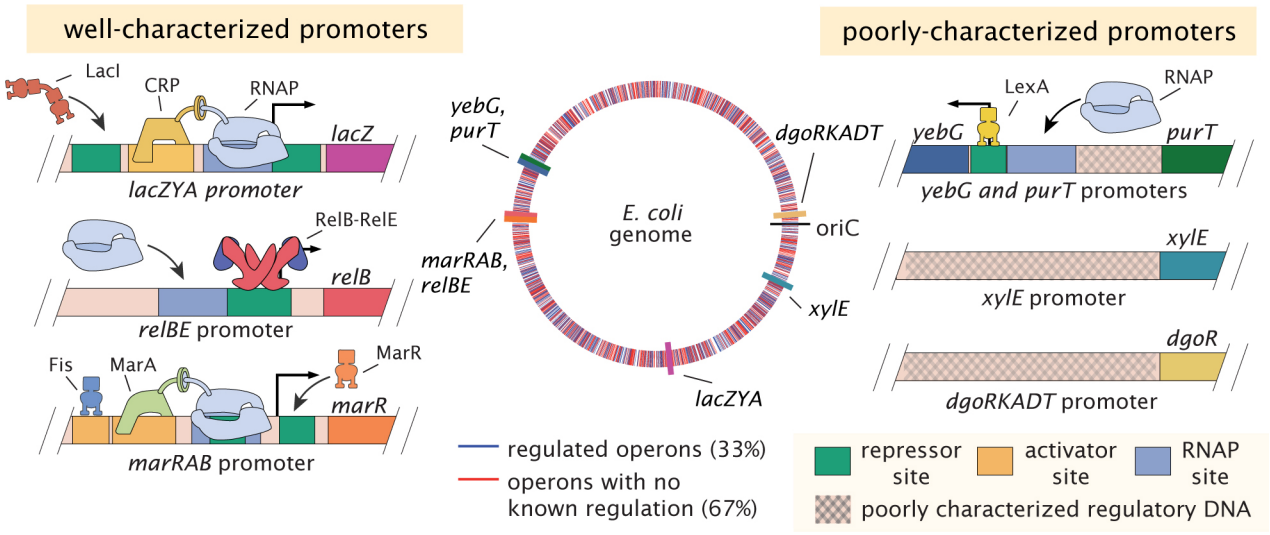

Figure 3.8: **Summary of transcriptional regulatory knowledge in *E. coli*.** left panel: Well-characterized promoters considered in this work. The schematics highlight the known regulatory architectures for the annotated promoters of *marRAB*, *relBE*, and *lacZYA*. The center plot identifies the genomic location of different operons in *E. coli*. Operons with annotated TF binding sites are shown in blue, while those lacking regulatory descriptions are shown in red [1]. The genomic location of the promoters considered in this work are labeled. Right panel: promoters associated with the operons of *yebG* and the poorly-characterized operons *purT*, *xylE*, and *dgoRKADT*. The promoters of *yebG* and *purT* are oriented in opposite directions. Repressor binding sites are shown in green, activator binding sites in yellow, and RNA polymerase (RNAP) binding sites in blue. The poorly characterized regulatory DNA is noted by a hashed pattern. The identification of regulated operons was performed using the annotated operons listed on RegulonDB [1], which are based on manually curated experimental and computational data. An operon was considered to be regulated if it had at least one transcription factor binding site associated with it.

protein showed a large differential expression level across growth conditions and suggests many of these unannotated promoters may possibly be under regulation.

Another way to view this data is to calculate the coefficient of variation (the ratio of the standard deviation to the mean protein copy number) for each gene across the 22 growth conditions. In Figure 3.9C, the coefficient of variation is plotted for each of the proteins measured in this study, separated by whether their promoter contains any known transcription factor binding sites (identified from RegulonDB [1]). For GalE, whose expression was perturbed by growth in galactose, we find a calculated coefficient of variation of 1.18. Using this as our reference for a regulated gene that was perturbed in the study, there appear to be many unannotated genes that may

in fact be under regulation. Among these, DgoD for example has a coefficient of variation of 3.64. Among the other proteins we investigated, XylE also has a high coefficient of variation, equal to 2.73, and shows almost no expression unless cells are grown in the presence of xylose as the carbon source. While we only pursued the promoters associated with expression of DgoR, DgoD, DgoK, DgoA, and XylE, there are many other unannotated promoters that will be of interest in future work.

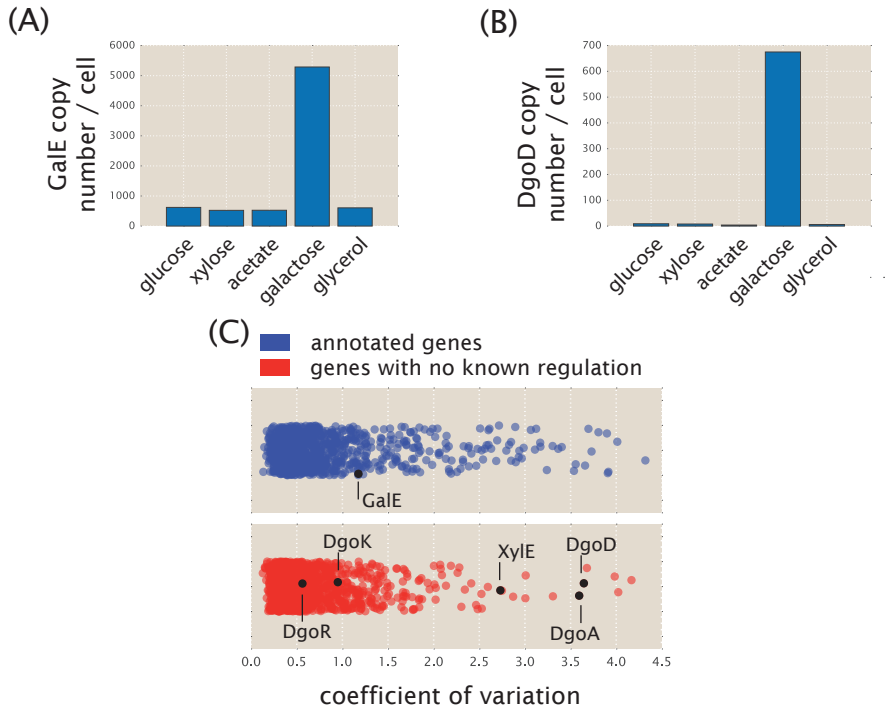

Figure 3.9: **Identification of unannotated genes with potential regulation and distribution of known transcription factor binding sites in *E. coli*.** (A) Here we show the protein copy numbers per cell for GalE across several carbon sources. Expression was sensitive to the presence of galactose which is consistent with its known regulation (with about 5000 copies per cell, versus about 500 for most other growth conditions). (B) DgoD was also found to be sensitive to the presence of galactose as the carbon source. The copy number was measured to be 675 copies per cell when cells were grown in galactose, and 15 copies per cell or less in all other conditions considered. For both (A) and (B), values are shown for growth in M9 minimal media, with glucose, xylose, acetate, galactose, and glycerol as carbon sources and obtained from [45]. (C) Coefficient of variation (standard deviation divided by mean copy number) across the 22 growth conditions for each protein measured in [45]. Proteins are identified as either having regulatory annotation (blue) or not (red) using the annotations in RegulonDB [1]. GalE is noted among the annotated genes and provides a reference as a gene that is known to be regulated and be perturbed in this study, as shown in (A).

3.6 Supplemental Information: Characterization of library diversity and sorting sensitivity.

Sort-Seq of the *rel* promoter using different sorting conditions.

In the work of the main text, Sort-Seq was performed by sorting cell libraries into four bins based on their fluorescence, each containing about 15 percent of the population. The remaining population was not collected and was discarded to waste. Due to the variability in expression of a single clonal population (Figure 3.6A), sorting into a larger number of narrower bins was not expected to provide additional resolution for the sequence-dependent fluorescence distribution. Given the success in identifying the known regulatory binding sites of the *lacZ*, *relB*, and *marR* promoters, and agreement between the inferred sequences logos and available sequence logos (see Figure 3.11), these conditions appeared to provide sufficient information to accurately analyze our libraries.

However, in order to further confirm that our results were not being influenced by the specific sorting scheme, we also tested several other sorting conditions using our *relB* promoter library. Here cells were sorted into either 4 or 8 bins, with a sorting gate containing between 10 and 22 percent of the population per bin. The associated expression shift plots and information footprints (defined in Supplemental Section 3.12) are shown in Figure 3.6B-D. In general we found little difference between each of these experiments. Energy matrices for the binding sites were similarly in agreement, with a Pearson correlation coefficient between matrix parameters generally greater than 0.9 across the different conditions tested.

Analysis of library diversity using data from the *mar* promoter.

Here we provide additional characterization of the mutagenized promoter libraries, using a library from the *marR* promoter as a representative example (70 bp region containing RNAP and MarR repressor sites). With the exception of the *lacZ* promoter, all library oligonucleotide pools were purchased from Integrated DNA Technologies (USA) with a target mutation rate of nine percent per nucleotide position. For the *lacZ* promoter library, we purchased an oligonucleotide pool using their Ultramer branded technology to allow for a longer mutagenized region that covered the known set of regulatory binding sites. While we intended to have a similar mutation rate, we found a mutation rate closer to three percent per nucleotide position. While unexpected, this allowed us to test two different mutation rates in our initial validation of the methodology using well-characterized promoters.

To get a better sense of how the mutation rate varies across the libraries, we plot a histogram of the number of mutations per base pair for the entire set of sequences found in the *marR* promoter library (Figure 3.6E). While we obtained an average mutation rate of 10.4% in this library, close to our target rate of 9%, there is some variability in this mutation rate as might be expected given that the incorporation of mutations in the DNA synthesis procedure is a random process. Since we are using these sequence data sets to infer sequence-specific models of binding between DNA and transcription factors, it was also of interest to consider the mutational coverage found within the library. As shown in Figure 3.6F, all single-point mutations and a large fraction of two-point mutations were present within the library. Due to the large number of possible three point mutants in a 60 bp region, only a small subset of the possible sequences will be found in the library.

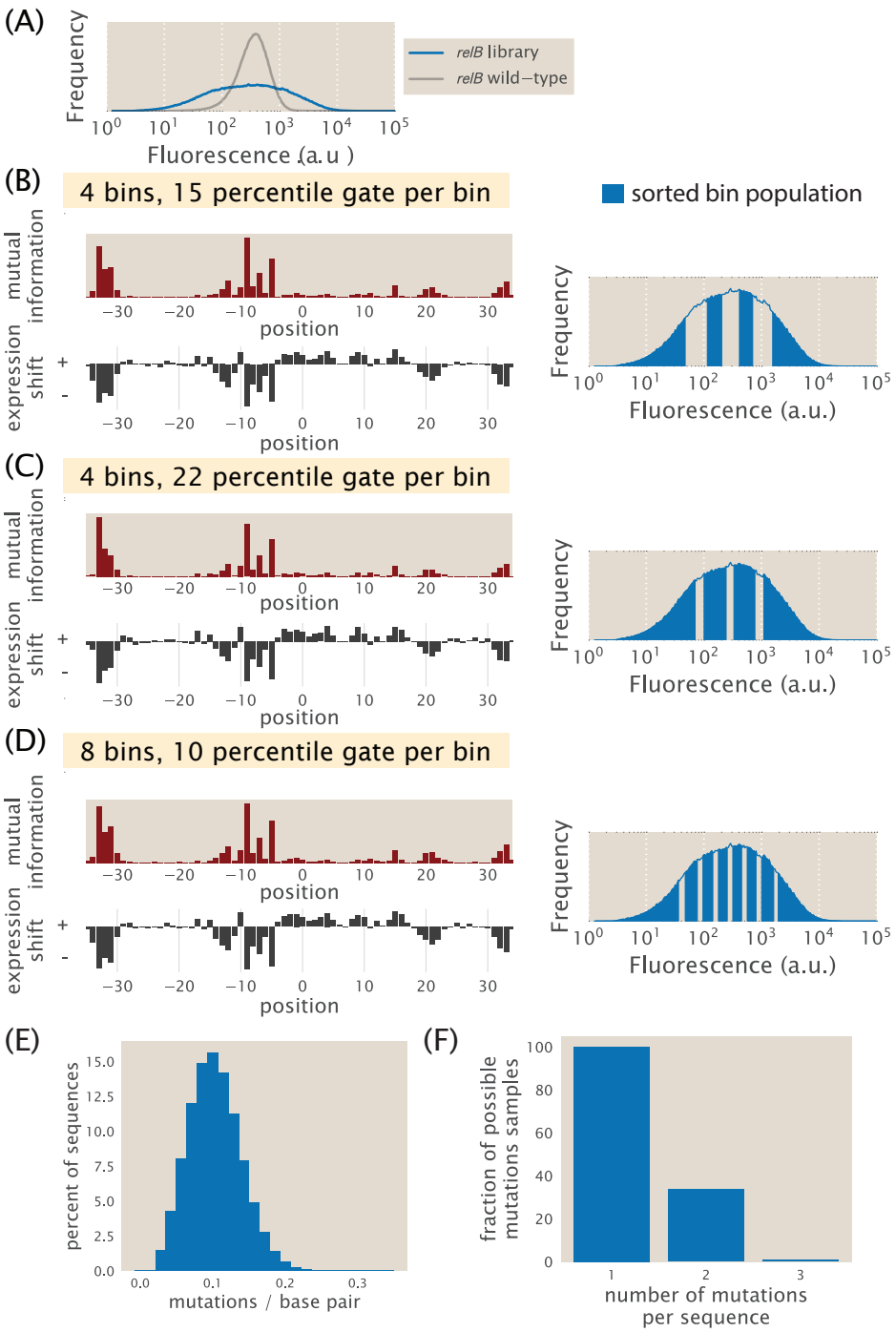

Figure 3.10: **Analysis of the library mutation spectrum and effect of Sort-Seq sorting conditions**.
(A) Here we used our *relBE* promoter library to test whether the sorting procedure influenced our Sort-Seq data analysis. The fluorescence histogram of the wild-type promoter plasmid (single clonal population) and the mutated library for the *relB* promoter are shown. Expression shifts and information footprints are shown for cells sorted under three different scenarios in (B) -(D). In (B) cells were sorted using the approach of the main text where cells were sorted into 4 bins, each containing 15% of the population. *(continued on next page)*

Figure 3.10: *(continued from previous page)* In (C) cells were similarly sorted into 4 bins, but where each bin contained about 22% of the population. In (D) cells were sorted into 8 bins, each containing about 10% of the population. The histograms beside each information footprint identify the approximate gating windows used to sort each fluorescence bin population. Histograms were based on between 400,000-500,000 cell counts. The same cell culture was used for each of the three Sort-Seq experiments performed here, sorted during the same sorting session. Cells were grown in M9 minimal media with 0.5% glucose like in the main text. (E) Histogram showing the mutation rate across all sequences found in the 60 bp *marRAB* library containing the RNAP and MarR repressor binding sites. Analysis was based on sequences from all fluorescence sorted bins. (F) The fraction of all possible unique sequences with one, two, or three mutations is shown for the *marRAB* library of (E). The coverage quickly drops for possible three-point mutations due to the large sequence space at this mutation frequency.

3.7 Supplemental Information: Generation of sequence logos.

Sequence logos provide a simple way to visualize the sequence specificity of a transcription factor to DNA, as well as the amount of information present at each position [26]. Here we describe how we generate them using either known genomic binding sites or the energy matrices that were determined from our Sort-Seq data. In each case we need to calculate a 4xL position weight matrix for a binding site of length L, which is used to estimate the position-dependent information content needed to construct a sequence logo.

Generating position weight matrices from known genomic binding sites.

From RegulonDB, we find there are $N_g = 260$ known binding sites for CRP on the *E. coli* genome [1]. To construct a position weight matrix using these genomic binding sites, we must first align all the sequences and determine the nucleotide statistics at each position. Specifically, we count the number of each nucleotide, N_{ij}, at each position along the binding site. Here the subscript i refers to the position, while j refers to the nucleotide, A, C, G, or T. We can then calculate a position probability matrix (also 4xL) where each entry is found by dividing these counts by the total number of sequences in our alignment,

$$p_{ij} = \frac{N_{ij}}{N_g}. \tag{3.1}$$

Note that in situations where the number of aligned sequences is small (e.g. less than five), pseudocounts [84] are often added to regularize the probabilities of the counts in the calculation of position probabilities,

$$p_{ij} = \frac{N_{i,j} + B_p}{N_g + 4 \cdot B_p}, \tag{3.2}$$

where B_p is the value of the pseudocount. The argument for their use is that when selecting from a small number of binding site sequences, just by chance infrequent nucleotides will be absent, and assigning them a probability (p_{ij}, noted above) of zero may be too stringent of a penalty [84, 85]. We let $B_p = 0.1$. In the limit of zero binding site sequences (i.e., with no sequences observed), this will result in probabilities p_{ij} approximately equal to the background probability used in calculating the position weight matrix below (and a non-informative sequence logo).

Finally, the values of the position weight matrix are found by calculating the log probabilities relative to a background model [86],

$$PWM_{ij} = \log_2 \frac{p_{ij}}{b_j}. \tag{3.3}$$

The background model reflects assumptions about the genomic background of the system under investigation. For instance, in many cases it may be reasonable to assume each base is equally likely to occur. Given that we know the base frequencies for *E. coli*, we choose a background model that reflects these frequencies (b_j: A = 0.246, C = 0.254, G = 0.254, and T = 0.246 for strain MG1655; BioNumbers ID 100528, `http://bionumbers.hms.harvard.edu`). From Equation 3.3, we can see that the value at the i, j^{th} position will be zero if the probability, p_{ij}, matches that of the background model, but non-zero otherwise. This reflects the fact that base frequencies matching the background model tell us nothing about the binding preferences of the transcription factor, while deviation from this background frequency indicates sequence specificity.

Generating position weight matrices from Sort-Seq data.

Next we construct a position weight matrix using the CRP energy matrix from our Sort-Seq data. Here we appeal to the result from Berg and von Hippel, that the logarithms of the base frequencies above should be proportional to their binding energy contributions [86, 87]. Berg and von Hippel considered a statistical mechanical system containing L independent binding site positions, with the choice of nucleotide b_j at each position corresponding to a change in the energy level by ε_{ij} relative to the lowest energy state at that position. This ε_{ij} corresponds to the energy entry in our energy matrix, scaled to absolute units, $A \cdot \theta_{ij} + B$ (where θ_{ij} is the i, j^{th} entry as noted in Supplemental Section 3.12). An important assumption is that all nucleotide sequences that provide an equivalent binding energy must have equal probability of being present as a binding site. In this way, we can relate the binding energies considered here to the statistical distribution of binding sites in the previous section. The probability p_{ij} of choosing nucleotide b_j at position i for protein binding will then be proportional to the probability that position i has energy ε_{ij}. Specifically, the probabilities will be given by their Boltzmann factors

normalized by the sum of states for all nucleotides,

$$p_{ij} = \frac{b_j \cdot e^{-\beta A \cdot \theta_{ij} \cdot s_{ij}}}{\sum_{j=A}^{T} b_j \cdot e^{-\beta A \cdot \theta_{ij} \cdot s_{ij}}}, \tag{3.4}$$

where $\beta = 1/k_B T$, with k_B is Boltzmann's constant and T the absolute temperature. Note that the energy scaling factor B drops out of this equation since it is shared across each term. As above, b_j refers to the background probabilities of each nucleotide.

One difficulty that arises when we use energy matrices that are not in absolute energy units is that we are left with an unknown scale factor A, preventing calculation of p_{ij}. We appeal to the expectation that mismatches usually involve an energy cost of 1-3 $k_B T$ [75]. In other work within our group, we have found this to be a reasonable assumption for LacI. Therefore, we approximate it such that the average cost of a mutation $\langle A \times \theta_{i,j} \rangle = 2k_B T$. We can then calculate a position weight matrix from Equation 3.3.

Construction of sequence logo

With our position weight matrices in hand we can now construct sequence logos by calculating the average information content at each position along the binding site. With our four letter alphabet there is a maximum amount of information of 2 bits ($\log_2 4 = 2$ bits) at each position i. The information content will be zero at a position when the nucleotide frequencies match the genomic background, and will have a maximum of 2 bits only if a specific nucleotide is completely conserved. The total information content at position i is determined through calculation of the Shannon entropy, and is given by

$$I_i = \sum_{j=A}^{T} p_{ij} \cdot log_2 \frac{p_{ij}}{b_i} = \sum_{j=A}^{T} p_{ij} \cdot \mathrm{PWM}_{ij}. \tag{3.5}$$

Here, PWM_{ij} refers to the i, j^{th} entry in the position weight matrix [86, 88]. The total information content contained in the position weight matrix is then the sum of information content across the length of the binding site.

To construct a sequence logo, the height of each letter at each position i is determined by

$$\mathrm{Seqlogo}_{ij} = p_{ij} \cdot I_i, \tag{3.6}$$

which is in units of bits. This causes each nucleotide in the sequence logo to be displayed as the proportion of the nucleotide expected at that position scaled by the amount of information contained at that position [26]. To construct sequence logos we use custom Python code written by Justin Kinney and available on our GitHub repository for this work (`https://www.github.com/RPGroup-PBoC/sortseq_belliveau`; DOI: `https://doi.org/10.5281/zenodo.1184169`).

Comparison of Sort-Seq sequence logos.

For the various annotated binding sites identified in this work we used our Sort-Seq data to generate energy matrices. While these energy matrices provide a concrete way to understand the sequence-dependent DNA-protein interaction, it was also useful to generate sequence logos from energy matrices to visually compare with sequence logos more conventionally generated using known genomic binding site sequences. In Figure 3.11 we show this comparison for transcription factors with three or more known genomic binding sites, with agreement more apparent when genomic binding site logos are based on a larger number of known sequences.

We also report the Pearson correlation coefficient between the position weight matrices from the Sort-Seq inference and the genomic alignment. To compare the two position weight matrices we first apply gauge fixing to each matrix in a similar manner as our energy matrix (see Supplemental Section 3.12). Each column is set to have a mean energy of zero and the matrix norm (or inner product) is normalized to have value one. Under this constraint, the Pearson correlation coefficient is simply given by the summed product of matrix entries,

$$r = \frac{COV(\text{PWM'}_X, \text{PWM'}_Y)}{\sigma_X \cdot \sigma_Y} = \sum_{i=1}^{L} \sum_{j=A}^{T} \text{PWM'}_{X,i,j} \cdot \text{PWM'}_{Y,i,j}. \tag{3.7}$$

Here, COV refers to the covariance between PWM'_X and PWM'_Y, where the superscript prime indicates that the matrices have been gauge fixed (mean energy in each column of zero and the matrix norm of 1). The subscript X, for example, would correspond to the Sort-Seq matrix, and Y, to the genomic matrix. σ_X and σ_Y refer to the standard deviation of the matrix entries for PWM'_X and PWM'_Y. We note that while Pearson correlation coefficient provide one useful metric to compare energy matrices, there are alternative metric that are also commonly used (Kullback-Leibler divergence, Euclidean distance, and Pearson χ^2 test, among others; see Gupta et al. 2007 [89], which is the publication for the TOMTOM motif comparison software and provides a good summary of these).

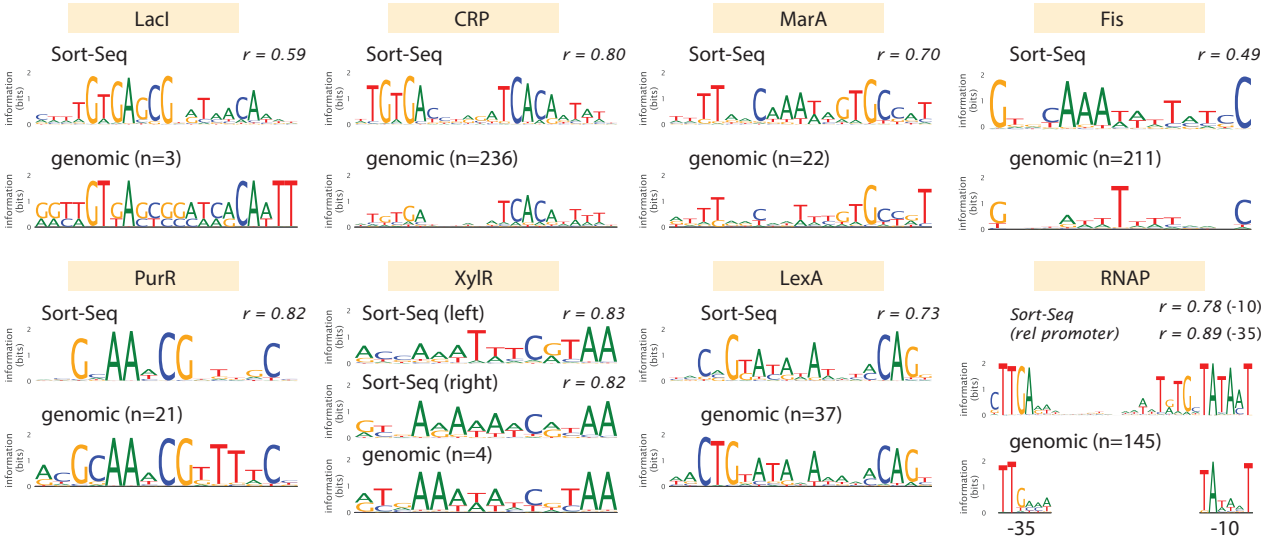

Figure 3.11: **Comparison between Sort-Seq and genomic-based sequence logos.** Comparisons are shown for LacI, CRP, MarA, Fis, PurR, XylR, LexA, and RNAP. Binding site sequences were obtained from RegulonDB, where *n* identifies the number of genomic binding sites that were used to construct the sequence logo. The Sort-Seq RNAP logo is based on data from the *rel* promoter. For the genomic RNAP logo, sequences were taken from computationally predicted RNAP binding sites on RegulonDB (top 3.3 % scored sequences using their reported metric) for the 6 bp regions of the -10 and -35 binding sites. Pearson correlation coefficients are calculated with Equation 3.7 using the position weight matrices from the Sort-Seq and genomic matrices. For LexA, the first four bp were not used in the calculation due to overlap with the -10 RNAP binding site of the *yebG* promoter.

3.8 Supplemental Information: Statistical mechanical model of the DNA affinity chromatography approach.

In order to better understand the factors that govern the success of the DNA affinity chromatography method, we took a statistical-mechanical approach to help identify the key parameters that will influence the fold enrichment of transcription factors that we measure. We are interested in calculating the probability that the transcription factor of interest binds to the target DNA sequence used for purification. We will ignore possible binding by proteins to the magnetic beads, to which the DNA oligonucleotides are tethered.

To calculate the probability that the transcription factor of interest is bound, we will simplify our problem by assuming that all other proteins in the lysate will bind the DNA with some average nonspecific binding energy. This must be included since these proteins will act as potential competitors for the tethered DNA. We must first enumerate the possible states of our DNA. For each DNA affinity purification, this will include the following three states: 1) no protein bound to the DNA, 2) the target transcription factor bound, and 3) a nonspecific protein is bound. These are shown in Supplemental Figure 3.12D for each of the DNA oligonucleotides used for the two different purifications performed.

The non-normalized probability of each state occurring is simply given by $e^{-\beta(\varepsilon_i - \mu_i)}$. Here, ε_i is the protein-DNA binding energy and μ_i, the chemical potential, for species i [90]. $\beta = 1/k_B T$, where k_B is Boltzmann's constant and T is the absolute temperature. The chemical potential contains information about concentration, and it is possible to alternatively write the non-normalized probability in terms of these, which is given by $C_i/C_o e^{-\beta\Delta\varepsilon_i}$. Here, C_i is the concentration of protein species i, and C_o, is the standard concentration, which is taken as 1 M. $\Delta\varepsilon_i$ is the binding energy for species i, relative to the unbound state.

We can now write the statistical weight for each state, which is summarized in Figure 3.12D. We allow the unbound state to act as our reference state with an energy equal to zero, and a corresponding statistical weight of 1. The probability of our target protein being bound to a certain DNA target, $P_{bound,DNA}$, will then be given by the statistical weight for the state where the target protein is bound, divided by the sum of statistical weights for each state. This is given by

$$P_{bound,DNA} = \frac{\frac{C_{TF}}{C_o} e^{-\beta\Delta\varepsilon_{TF,DNA}}}{1 + \frac{C_{ns}}{C_o} e^{-\beta\Delta\varepsilon_{ns}} + \frac{C_{TF}}{C_o} e^{-\beta\Delta\varepsilon_{TF,DNA}}}, \tag{3.8}$$

where the subscript 'TF, DNA' identifies the target transcription factor and its binding to a specific DNA target. In regard to our two purifications shown in Figure 3.12D, $\Delta\varepsilon_{TF,s}$ refers to the binding energy of the transcription factor to its target binding site, while $\Delta\varepsilon_{TF,ns}$ refers to the nonspecific binding energy to non-target reference DNA. In addition, $\Delta\varepsilon_{ns}$ refers to the binding energy of other proteins present in the lysate, which may bind the DNA nonspecifically.

We can now calculate the fraction of bound transcription factor, $P_{bound,DNA}$, using some reasonable values for *E. coli* [51, 91]. Here we use $C_{TF} = 10^{-8}M$ (about 10 copies per cell), $C_o = 1M$, $\Delta\varepsilon_{TF,s}$ = - $15k_BT$, and $\Delta\varepsilon_{ns}$ = - $5k_BT$. $C_{ns} = 3 \cdot 10^{-3}M$, which is the approximate number of proteins in *E. coli*. The specific numbers will depend on the DNA target sequence used, the concentration of target protein, as well as the lysate preparation itself. Here we find $P_{bound} \approx 0.02$. In contrast, for the nonspecifically bound fraction we calculate about a ten-fold higher fraction of protein bound to the DNA. Even though the binding energy for a target transcription factor is significantly stronger than the competitor proteins that bind nonspecifically, the target transcription factor is generally several orders of magnitude lower in abundance. This result in particular highlights our rationale for using a additional reference purification to distinguish the target transcription factor from non-specifically bound proteins [21]. We consider the consequences of this next.

In this second reference purification, the DNA no longer has the target binding site, and thus the value of $P_{bound,DNA}$ for the transcription factor should be significantly smaller. We can use Equation 3.8 to calculate expected ratio of transcription factor bound to target DNA versus reference DNA, given by

$$\frac{P_{bound,target}}{P_{bound,reference}} = \frac{\frac{C_{TF}}{C_o}e^{-\beta\Delta\varepsilon_{TF,s}}}{1 + \frac{C_{ns}}{C_o}e^{-\beta\Delta\varepsilon_{ns}} + \frac{C_{TF}}{C_o}e^{-\beta\Delta\varepsilon_{TF,s}}} \cdot \frac{1 + \frac{C_{ns}}{C_o}e^{-\beta\Delta\varepsilon_{ns}} + \frac{C_{TF}}{C_o}e^{-\beta\Delta\varepsilon_{TF,ns}}}{\frac{C_{TF}}{C_o}e^{-\beta\Delta\varepsilon_{TF,ns}}} \tag{3.9}$$

$$= \frac{e^{-\beta\Delta\varepsilon_{TF,s}}}{e^{-\beta\Delta\varepsilon_{TF,ns}}} \frac{1 + \frac{C_{ns}}{C_o}e^{-\beta\Delta\varepsilon_{ns}} + \frac{C_{TF}}{C_o}e^{-\beta\Delta\varepsilon_{TF,ns}}}{1 + \frac{C_{ns}}{C_o}e^{-\beta\Delta\varepsilon_{ns}} + \frac{C_{TF}}{C_o}e^{-\beta\Delta\varepsilon_{TF,s}}}. \tag{3.10}$$

Again, the subscript $\Delta\varepsilon_{TF,ns}$ refers to the binding energy of the transcription factor to the non-target (i.e. non-specific) reference DNA. Using the example values from our

calculation of P_{bound} above, we find that $1 + \frac{C_{ns}}{C_o} e^{-\beta \Delta \varepsilon_{ns}} \gg e^{-\beta \Delta \varepsilon_{TF,s}} \gg e^{-\beta \Delta \varepsilon_{TF,ns}}$, with Equation 3.10 simplifying to

$$\frac{P_{bound,target}}{P_{bound,reference}} \approx \frac{e^{-\beta \Delta \varepsilon_{TF,s}}}{e^{-\beta \Delta \varepsilon_{TF,ns}}} = e^{-\beta(\Delta \varepsilon_{TF,s} - \Delta \varepsilon_{TF,ns})}. \tag{3.11}$$

This result suggests that the enrichment ratio should mainly depend on the difference in binding energy between the DNA sequences used in the two purifications. Our results from purifying LacI with strains containing different LacI copy number per cell and with different DNA target sequences (see Figure 3.12C) appear to agree with this result in general, where we see greater enrichment when using the strong Oid target LacI binding site sequence than the weaker O3 binding site sequence. This appears to influence the enrichment ratio more significantly than protein concentration, although further work will be needed to fully characterize this relationship.

3.9 Supplemental Information: DNA affinity chromatography and mass spectrometry experimentation and analysis.

In this section we provide additional details on the use of DNA affinity chromatography and mass spectrometry to identify the transcription factors that bind to our putative binding sites. In particular, we provide additional data to demonstrate protein labeling and characterize the dynamic range expected from our enrichment measurements (see Methods Section for more details about the approach). We also provide data from an affinity chromatography experiment in which the same DNA oligonucleotide sequence was used for both target and control purifications. The ideal result from such an experiment is that each protein detected is found in equal abundance between the two purifications performed, yielding an enrichment ratio equal to one. However, there is some inherent variability in such a measurement and we provide some characterization of that uncertainty here. Lastly, we provide additional data showing that we can purify and identify transcription factors at concentrations ranging from about 10 to 1,000 copies per cell.

Characterization of SILAC labeling and measurement of protein enrichment ratios.

To ensure *E. coli* cells incorporated the heavy isotope of lysine ($^{13}C_6^{15}N_2$-L-lysine, heavy lysine), we first generated an auxotropic strain which was unable to synthesize its own lysine through deletion of the *lysA* gene [92]. LysA is an enzyme that catalyzes the last step in lysine biosynthesis. Furthermore, to ensure proteins would be sufficiently labeled when growing cultures for lysate preparation we inoculated our cultures with a large dilution of 1:5,000. This large dilution is important since the inoculate represents an unlabeled fraction of the cell population. We checked the effective labeling efficiency by combining lysates from cells grown with heavy and light (natural) lysine over a range of ratios between 0.1/1 to 1,000/1 (heavy / light). The measured ratio in abundance for each of the proteins detected among the two lysates are plotted in Figure 3.12A. In calculating these values, we found that the median average was measured to be 0.71 (heavy / light). We do not expect a discrepancy between measured heavy and light protein of similar abundance, and this suggested there may have been some inaccuracy in the Bradford assay used to measure protein concentration prior to mixing our lysates. We therefore renormalized the ratios according to this measured ratio. The data suggests a labeling efficiency of at least 99% (red dashed line, in comparison to perfect labeling shown by the gray dashed line). One important aspect highlighted by this data is

that the highest enrichment ratio we should expect to measure in our DNA affinity experiments is several hundred fold.

Characterization of protein enrichment variability from identical DNA targets. For each DNA affinity chromatography experiment, we are trying to identify a DNA-binding protein that shows up in higher abundance when we use the target binding site sequence identified by Sort-Seq (i.e. a transcription factor binding site), relative to a purification where that target sequence has been mutated away. To ensure that our measured enrichment ratios were not an artifact of noise in the measurement, it was important to also check the measurement variability when both lysate purifications used an identical DNA sequence. In this way, we could characterize the inherent variability in such a measurement. To proceed, we performed experiments using the control DNA sequence that was used in our purification of the *purT* promoter target (Fig. 5C, though any DNA oligonucleotide could have been used). We performed this in triplicate and consider the average enrichment ratios for each protein measured across the three experiments. In the left panel of Figure 3.12B we show the average enrichment values that were measured for each of the detected proteins. Since many of the data points fall on top of one another, we also provide a histogram of the associated data (Figure 3.12B, right plot). Here we have taken the logarithm of the enrichment ratios so that the bins are equally spaced. The shaded region in both plots identifies the range between the 2.5^{th} and 97.5^{th} percentiles, highlighting that the majority of proteins were found between an enrichment ratio of 0.2 and 3.3 (or log enrichment ratio of between -1.5 and 1.2). The ideal enrichment expected would be a value of 1.0 or log ratio of 0. In the main text, the enrichment values for transcription factors found using targets associated with the *lacZ*, *relB*, *purT*, *xylE*, and *dgoR* promoters fall well outside of the range of variability established here.

Identification of LacI by mass spectrometry using strains with a variable LacI copy number.

Finally, one experiment that we performed, in addition to purifying LacI with different strength binding site targets (i.e., Fig. 4A), was to consider the copy number per cell of the LacI target, as copy number should influence the fraction of bound LacI (see details in Supplemental Section 3.8). Here we used strains whose protein concentration has been measured during growth in M9 minimal media with 0.5% glucose and whose average LacI number had previously been measured to range from the native expression of 11 ± 2 tetramers per cell, to a maximum concentration

of 870 ± 170 tetramers per cell. In Figure 3.12C we show the enrichment ratios measured for LacI from individual experiments (n = 1-2 per strain). Here we were able to purify LacI using either the weak O3 or strong Oid binding site sequence for each of the different strains, though we also see that the O3 target sequence provides an enrichment that is much closer to the tail of the control experiment in Figure 3.12B. Additionally, while the copy number of LacI appears to affect the enrichment ratio in some experiments, it does not have a consistently significant effect.

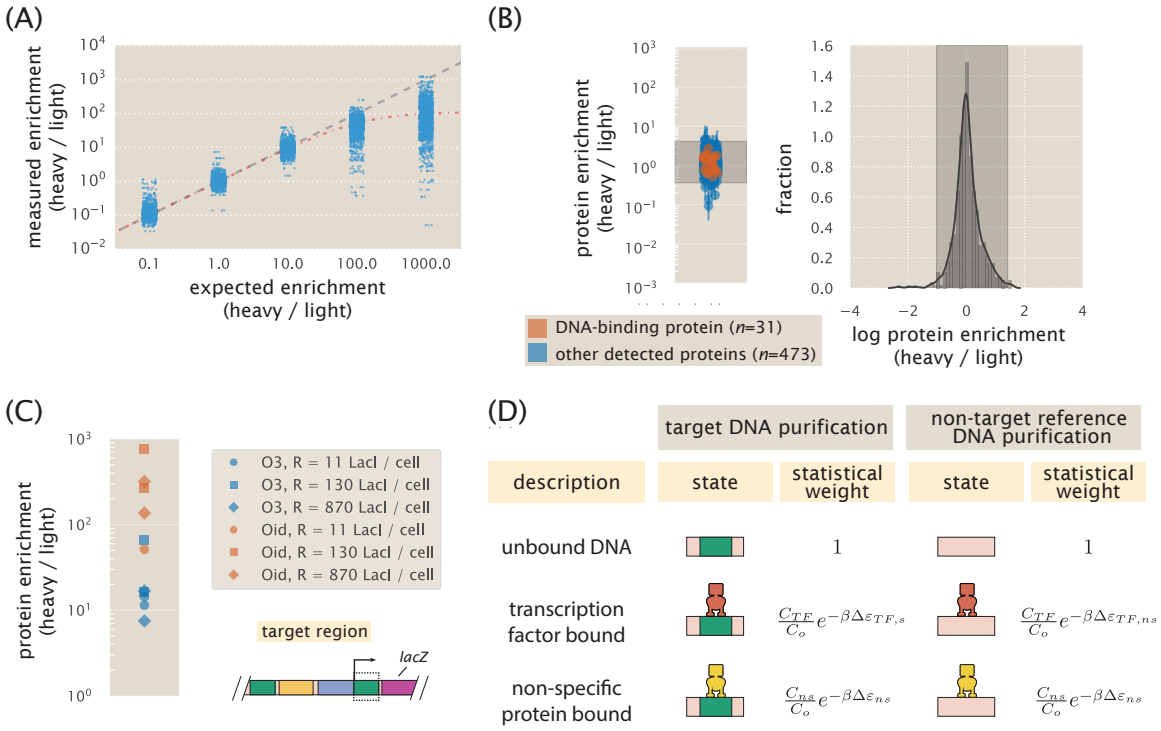

Figure 3.12: **Identification of transcription factors using DNA-affinity chromatography and mass spectrometry.** (A) Characterization of stable isotopic lysine labeling and mass spectrometry measurement sensitivity. Lysates from cell cultures grown in either heavy ($^{13}C_6^{15}N_2$-L-lysine) or normal L-lysine were combined at ratios between 0.1:1 to 1000:1 heavy:light and the measured ratios in abundance are plotted for each protein. Note that for the 1:1 ratio we found a median ratio of 0.71. We therefore renormalized the ratio values using this as a correction factor. Data points represent the average values from $n = 3$ replicates. The gray line represents the expected measurement under perfect labeling, while the red line represents a 99.1 % labeling efficiency (assuming that some fraction of heavy lysate is unlabeled). (B) DNA-affinity purification using the same DNA oligonucleotide to purify protein for both heavy and light cell lysates ($n = 3$). The scatter plot shows the average enrichment values for each protein detected. Proteins with DNA binding motifs [2] are shown in red ($n = 41$), while other detected proteins are in blue ($n = 581$). Error bars represent the standard deviation, calculated from log protein enrichment values. The histogram shows the distribution of the measured ratios for all detected proteins, with 95% of the measurements contained between a log enrichment of -1.5 and 1.2, as indicated by the shaded region. Lysates were prepared from cells grown in M9 minimal media with 0.5% glucose. (C) DNA-affinity purification of LacI using three different *E. coli* strains with repressor copy numbers per cell of 11 ± 2, 130 ± 20, and 870 ± 170 (tetramers per cell) [34]. Operator strength was varied by purifying LacI with either the weak O3 or strong Oid operators. LacI was detected as the most significantly enriched protein among all proteins detected. Each data point represents the enrichment from a single purification experiment (n = 1-2 for each strain). (D) States and weights are shown for an oligonucleotide in which a target transcription factor and other cellular proteins compete for a DNA binding site. Within the cell lysate, the target protein is present at a concentration C_{TF}, while all other proteins, which may bind the DNA nonspecifically are present at a concentration C_{ns}. C_o is the standard concentration. The difference in energy between a repressor bound to the target DNA binding site and an unbound DNA is $\Delta\varepsilon_{TF,s}$ when the binding site is present. Otherwise, the binding energy is given by $\Delta\varepsilon_{TF,ns}$. Other proteins that bind nonspecifically, irrespective of the DNA sequence, have a binding energy of $\Delta\varepsilon_{ns}$.

3.10 Supplemental Information: Selection of the mutagenesis window for promoter dissection by Sort-Seq.

In designing our mutagenized promoter libraries, we found it useful to consider what was known regarding both the genes of interest and general patterns of transcriptional regulation in *E. coli* and bacteria more broadly. Two useful resources were RegulonDB [1] and EcoCyc [2], which summarize much of what is known about transcriptional regulation in *E. coli*. RegulonDB, in particular, aims to compile all available data regarding gene regulation in *E. coli* into a single database and is the most complete record available for *E. coli* [93].

While Sort-Seq enables us to identify all proteins involved at a promoter, one potential limitation is that a transcription factor binding site will only be identified if it was contained within our mutagenized region. Using the known transcription factor binding sites in *E. coli* as a guide in our design, we made an educated guess regarding where we should search for transcription factor binding sites. Figure 3.13 shows a histogram of all of the transcription factor binding site positions from RegulonDB. By staggering a set of 60bp windows to cover a 150 bp region, we found we would expect to capture 73 percent of the known transcription factor binding sites. We chose 60 bp-70 bp windows for most libraries since they could be readily synthesized by Integrated DNA Technologies (USA) and were more economical than longer oligonucleotides. We also included about 15 bp of overlap between staggered regions to provide some replicates of the mutated base pairs on the different libraries.

It is also useful to note that our approach does not require that this specific strategy be used to create mutagenized promoter constructs. The methodology only requires compatibility between the length of the mutagenized region probed and the sequencing platform used. Microarray synthesized oligonucleotides provide another approach for targeted oligonucleotide design [94], and error-prone PCR can enable longer mutagenized windows within a single library [59, 95]. In addition, advances in sequencing, either through longer reads or alternative sequencing platforms such as PacBio (Pacific Bioscience, USA) and MinION (Oxford Nanopore Technologies, UK) are making it possible to sequence longer mutagenized regions, and CRISPR technologies could make it possible to identify longer range interactions such as DNA looping in bacteria (e.g. the 1 megabase region considered in [20]).

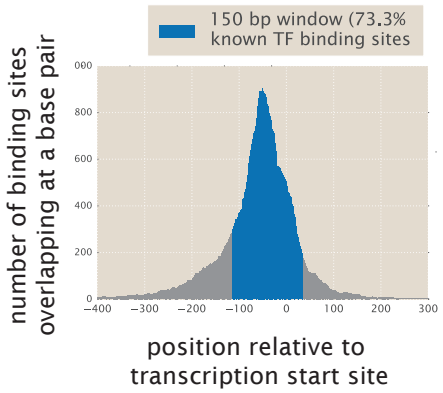

Figure 3.13: **Distribution of known transcription factor binding sites in *E. coli*.** The histogram shows the genome-wide distribution of transcription factor binding sites relative to their respective transcription start sites. Binding sites were compiled from RegulonDB and used to calculate the number of overlapping binding sites at each position using the length and position of each binding site sequence. The location of the 150 bp mutation window used in this study is shown in blue, expected to capture upwards of 70% of known transcription factor binding site position.

3.11 Supplemental Information: Additional data from Sort-Seq experiments of the main text.

Here we provide additional data and analysis on the promoters of *rel*, *mar*, *yebG*, *purT*, *xylE*, and *dgoR* to provide additional support for the results and conclusions made in the main text.

The *rel* and *mar* promoters

In our analysis of the *rel* and *mar* promoters in the main text, it was noted that the sequence specificity of the repressors RelBE and MarR lacked any prior characterization. In order to validate that the observed features of the expression shift plots were due to binding by these regulatory proteins, we performed additional Sort-Seq experiments in deletion strains for these regulators. The expression shift plots were shown in the main text (Fig. 3). Here we provide a more quantitative analysis to show that the energy matrices for binding by RelBE and MarR poorly describe the sequence data when *relBE* and *marR* are deleted, respectively.

Since the transcription factors have been deleted, we expect the energy matrix predictions of each sequence's binding energy to provide no clear trend across the sorted bins (i.e., zero or little mutual information). To first give a sense for how mutual information is calculated, in Figure 3.14A and Figure 3.14B we first show the estimated joint distributions when we apply the RelBE energy matrix (from Fig. 2B of main text) to either a replicate Sort-Seq experiment or to the $\Delta relBE$ deletion data. When applying the RelBE energy matrix to the wild-type data, we find a clear trend, with the strongest binding energies (lowest rank order) more likely found at the lowest fluorescence bin, and the weakest binding energies more likely found in the highest fluorescence bin.

Next we focus in on our data from the deletion strains of $relBE$ and $marR$ (Figure 3.14C and 3.14D, respectively). In each case, we find that our energy matrices poorly describe the data and are not substantially better than a randomly generated matrix. In Figure 3.14B it might have been noted that there were still some positions with non-zero expression shift (i.e., still appear informative). In order to show that this remaining information cannot be accounted for from our energy matrices, we also estimated the maximum information present in the Δ strain data sets (by directly fitting a matrix to the Δ strain data). Importantly, we find that this remaining information cannot be explained by our RelBE and MarR energy matrices, and must be due to other features in the data (for example, for MarR, this region overlaps with

the region that RNAP binds).

The *yebG* promoter

The *yebG* promoter is among a variety of genes known to increase expression when cells are under DNA damage stress [49], and shared the intergenic region with the *purT* promoter. In the main text we considered the *yebG* promoter in cells grown in standard M9 minimal media with 0.5% glucose (Fig. 5A). While the expression shifts appeared to align with annotated binding sites for LexA (positive shift), and the RNAP binding site (negative shift), we did not show evidence for the identity of each binding protein in the main text. Here we present results from our inference of energy matrices using our Sort-Seq data, which confirm the identity of the binding proteins. We also explore the regulation of *yebG* by perturbing the regulatory state through induction of the SOS response [48, 49].

We begin by considering the Sort-Seq data from cells grown in M9 minimal media with 0.5% glucose. In Figure 3.15A we show the inferred energy matrices associated with the annotated site for LexA. This was in excellent agreement with the known sequence specificity of LexA (see Figure 3.11 for a direct comparison with the genomic sequence logos). We note, however, that the RNAP binding site was shifted by 9 bp from the annotated binding site [48], with an overlap between the -10 RNAP site and 4 bp of the LexA binding site.

We were also interested in confirming that the *yebG* promoter responds DNA stress and is induced as part of the SOS response. By repeating Sort-Seq in cells grown in non-lethal concentrations of mitomycin C (1 μg/ml) [48] we observed a dramatic increase in expression relative to growth without mitomycin C. Fluorescence histograms showing expression from our plasmid reporter in non-mutagenized promoter constructs are shown in Figure 3.15B. From the expression shift plots and information footprint (which are defined in Supplemental Section 3.12 and used in Kinney *et al.* [13]) in Figure 3.15D we find that this is due to loss of repression at the LexA binding site. This is consistent with the expectation that LexA undergoes proteolysis as part of the SOS response [49].

The *purT* promoter

When cells were grown in the presence of adenine, we identified a putative repressor site between the -10 and -35 regions of the RNAP binding site of the *purT* promoter. In our initial attempt to identify the associated transcription factor we performed a DNA affinity purification using conditions that matched the growth conditions where

repression was observed. However, as shown in Figure 3.15C, the most significantly enriched protein (GlpR) only showed an enrichment of about 2.9, which was near the shaded region associated with most other non-specific proteins detected. Only upon repeating our purification in the presence of hypoxanthine (10 μg/ml) (Fig. 5C) did we find enrichment of PurR (approximately 350 fold relative to our reference purification).

The *xylE* promoter

In the main text it was noted that we could not perform Sort-Seq on the *xylE* promoter unless cells were grown in xylose. In Figure 3.15E, we show the associated fluorescence histograms from libraries grown in either glucose or xylose. Interestingly, each mutated window was essentially identical to autofluorescence when cells were grown in glucose. In contrast, growth in xylose showed differential expression for each of the mutated regions. While the promoter was expected to be sensitive to the presence of xylose (causing an increase in expression [45]), this was still a non-obvious result without prior knowledge of whether repressors or activators were involved.

In our analysis we also noted that the identified set of activator binding sites conformed well with the two other promoters regulated by XylR and CRP, namely *xylFG* and *xylAB*. Here we scanned our inferred energy weight matrix across the intergenic regions of *xylFG* and *xylAB*, in order gain further confidence that the identified feature matched the known binding specificity of these transcription factors. These are shown in Figure 3.15F. At each position in these plots, we use the energy matrix to calculate the binding energy of the putative transcription factors. For each we identify a strong peak that does indeed align well with the annotated binding sites of XylR and CRP. While our predicted binding energies are not in absolute k_BT units, they are much more negative than the promoter background and predict a similar binding energy (in arbitrary units) to the binding site region of the *xylE* promoter.

The *dgoR* promoter

The last promoter we considered was associated with the expression of the *dgoRKADT* operon. Due to the complexity observed, we were unable to show all data in the main text that supported our identification of the regulatory architecture. In particular, here we show the sensitivity to the different carbon sources considered and additional analysis of the identified regulatory binding sites for DgoR, RNAP, and CRP.

The *dgoR* promoter is induced when cells are grown in galactose and D-galactonate.

Prior to performing Sort-Seq on this promoter, we confirmed prior observations that expression was sensitive to the presence of galactose and D-galactonate [45, 56]. Using a wild-type promoter plasmid for the *dgoR* promoter, cells were grown in M9 minimal media with either 0.5% glucose, 0.23% D-galactose, or 0.23% D-galactonate. Fluorescence histograms are shown in Figure 3.16A, where we observed higher expression in galactose over glucose, and even higher expression when cells were grown in D-galactonate.

An RNAP binding site is apparent in the downstream region of the *dgoR* promoter when cells were grown in glucose.

In Fig. 7A we showed plots comparing the expression shifts upon mutation when cells were grown in either glucose or D-galactonate. In Figure 3.16B we reproduce the expression shift plots along with an energy matrix for the region from approximately -70 to -30, which helped us to identify the RNAP binding site in this region. While the -10 TATAAT motif is quite apparent, the -35 site is less clear. Interestingly, while the -35 region shows a most energetically favorable sequence of TTTACA (close to the consensus of TTGACA), the wild-type sequence is CCCCCC and suggests this is a weak RNAP binding site.

Deletion of the *dgoR* gene recovers the induced phenotype.

Comparing the expression shift values at each position in cells grown in either glucose or D-galactonate, we find that they are poorly correlated (Figure 3.16C, left plot). However, upon identifying DgoR as a putative regulator in the upstream region of the promoter, we then performed Sort-Seq in a $\Delta dgoR$ strain. This was shown in Fig. 7D with cells grown in glucose. Interestingly, the expression shifts were much more similar to the wild-type cells grown in D-galactonate, suggesting that deletion of *dgoR* has switched regulation to the induced state (Figure 3.16C, right plot).

While it is unclear what causes the noisy profiles in the expression shift plots, one hypothesis was that the different RNAP binding sites were producing at least two distinct mRNA transcriptions, whose 5' untranslated might influence transcript stability and GFP expression. In particular, the upstream RNAP binding site will generate a much longer 5' untranslated region, and mutations that influence mRNA

structure and stability might show up as an effect on expression within the region we considered by Sort-Seq. Using the Salis lab ribosomal binding site calculator [96] and RNA structure predictions with NUPACK [97], we predicted the secondary structure of the two expected mRNAs transcripts (Figure 3.16D). We find that the longer transcript (expected when cells are grown with D-galactonate), does indeed predict a strong secondary structure that alter translation from this transcript.

Simulations of upstream promoter region identify multiple overlapping RNAP binding sites.

Next we consider additional analysis to support the presence of overlapping RNAP sites that was noted in Fig. 7C. Since Sort-Seq does not differentiate between multiple transcription start sites, the sorted data will represent a mixture of all transcripts generated from the promoter. Using our RNAP energy matrix from the *relBE* promoter (with an additional 1 bp spacer included to increase the distance between -10 and -35 to 18 bp), we were able to identify multiple overlapping sequences that each predicted a similar binding energy by RNAP. The sequence logo in Fig. 7C of the main text (top logo) therefore likely represents the convolution of these multiple binding sites and would explain why we do not see the conventional -35 RNAP motif in the sequence logo.

To convince ourselves that this was a reasonable hypothesis, we performed several Sort-Seq simulations of the *dgoR* promoter to estimate what we may have expected if 1-3 of these identified RNAP binding sites were functional. These simulations use energy matrices and a thermodynamic model of regulation to predict gene expression as a function of regulatory sequence in an attempt to mimic a real Sort-Seq experiment. The code used is available on our GitHub repository (`https://www.github.com/RPGroup-PBoC/sortseq_belliveau`; DOI: `https://doi.org/10.5281/zenodo.1184169`) and we briefly describe the approach here. We began by first generating a library of five million mutated *dgoR* promoter sequences (10% mutation rate). We then assumed that transcription from each RNAP is proportional to $P/N_{NS} \cdot e^{-\beta E}$ where P is the RNAP copy number per cell, $N_{NS} = 4.6 \times 10^6$ refers to the number of non-specific binding sites on the genome, and $\beta = 1/k_B T$, where k_B is Boltzmann's constant and T is the absolute temperature. We introduced noise into our simulation by assuming that the RNAP copy number P was normally distributed across our library with a mean value of $3,000$ and standard deviation of 750 copies per cell [45, 98]. As defined in Supplemental Section 3.12, E represents

the binding energy as determined from the energy matrix.

Using these calculations to predict expression from each mutated sequence, the sequences were then computationally sorted in the same manner as that performed experimentally. We did this assuming the presence of one, two, or three active RNAP binding sites based on those identified. As shown in Figure 3.16F, the presence of three RNAP binding sites produces a result that conforms much better with experimental results than the presence of only one RNAP binding site. Note that binding sites were successively included into the model based on their predicted binding energies (wild-type RNAP 1: -1.99 a.u., wild-type RNAP 2: -1.74 a.u., wild-type RNAP 3: -1.60 a.u.; versus an average of -0.14 a.u. and standard deviation of 0.56 a.u. when the energy matrix is scanned across the promoter).

The presence of the class II CRP activator binding site is enhanced using strain JK10, grown with cAMP.

Lastly, we show additional evidence to support the claim of a putative binding site for CRP. Since CRP binds to DNA by co-activation through binding with cAMP, we used the strain JK10 (based on TK310 [13]; MG1655 $\Delta cyaA\Delta cpdA$), where we could control binding of CRP to DNA by direct supplement of cAMP to the growth media. Here we grew cells in EZrich MOPS media (Teknova, CA, USA) with D-galactonate as the carbon source and supplemented with 500 μM cAMP. While the sequence logos in Fig. 7E showed a good match with the left site of the CRP binding site, our hypothesis here was that addition of a high concentration of cAMP might enhance the CRP motif in our data. This appeared to be the case, and the right side of the binding site (which overlaps the -35 RNAP binding site) shows a stronger preference for the sequence CAC than present with the wild-type *E. coli* strain (important for binding by CRP in both the *lac* and *xylE* promoters).

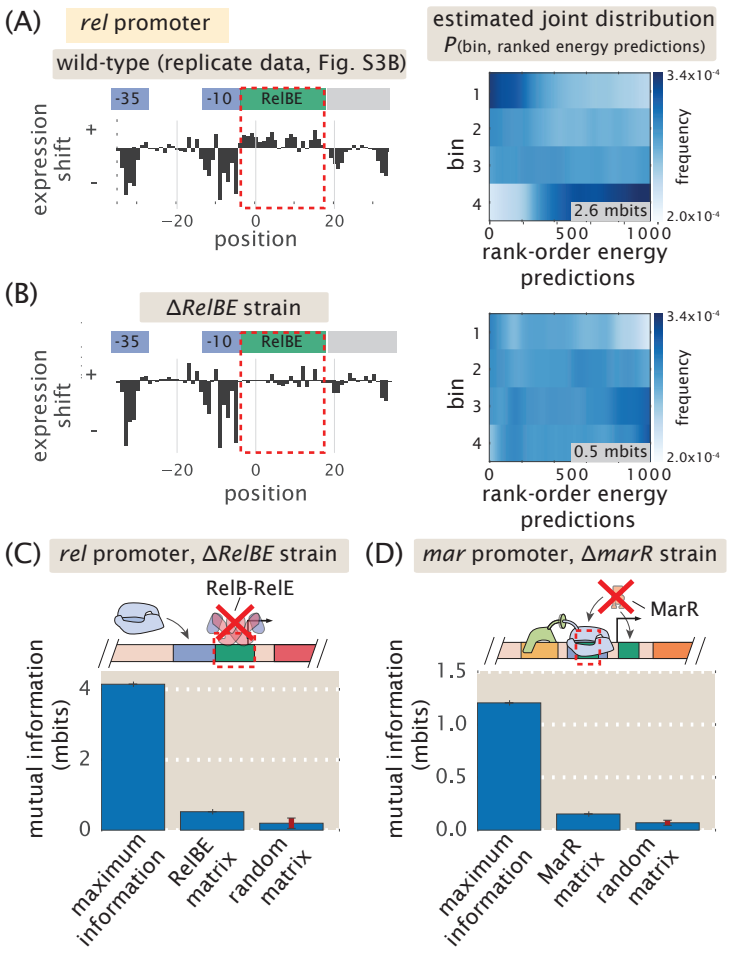

Figure 3.14: **Predictive information of transcription factor energy matrices when applied to Sort-Seq data.** In (A) and (B) we use our RelBE energy matrix to predict binding energies across all sequence data for a replicate experiment with wild-type *E. coli* and a $\Delta relBE$ strain, respectively. The 2-d histograms show the estimated joint probability distributions between bin and rank-ordered energies (generated by binning sequences into 1000 bins). The calculated information (in mbits) shown in the joint distribution plot represents the mutual information from these estimated joint distributions. In (C) and (D) we focus on our transcription factor deletion strains ($relBE$ in (C) and $marR$ in (D)), and similarly calculate mutual information between bin and energy matrix predictions (again, using their rank-ordered predictions). The 'maximum information' represents the estimated maximum information that might be obtained by fitting an energy matrix to the Δ strain data. The 'random matrix' represents the average mutual information calculated from 20 randomly generated energy matrices (error bar represents standard deviation) applied to the sequence data. To provide consistent comparisons, all matrices were 'gauge fixed' such that the mean energy in each column of zero and the matrix norm of 1. Note that for MarR we show analysis for the left MarR binding site. In the right binding site, there is additional information corresponding to the ribosomal binding site. The joint probability distribution and associated mutual information are calculated following the procedure described in Section 3.12.

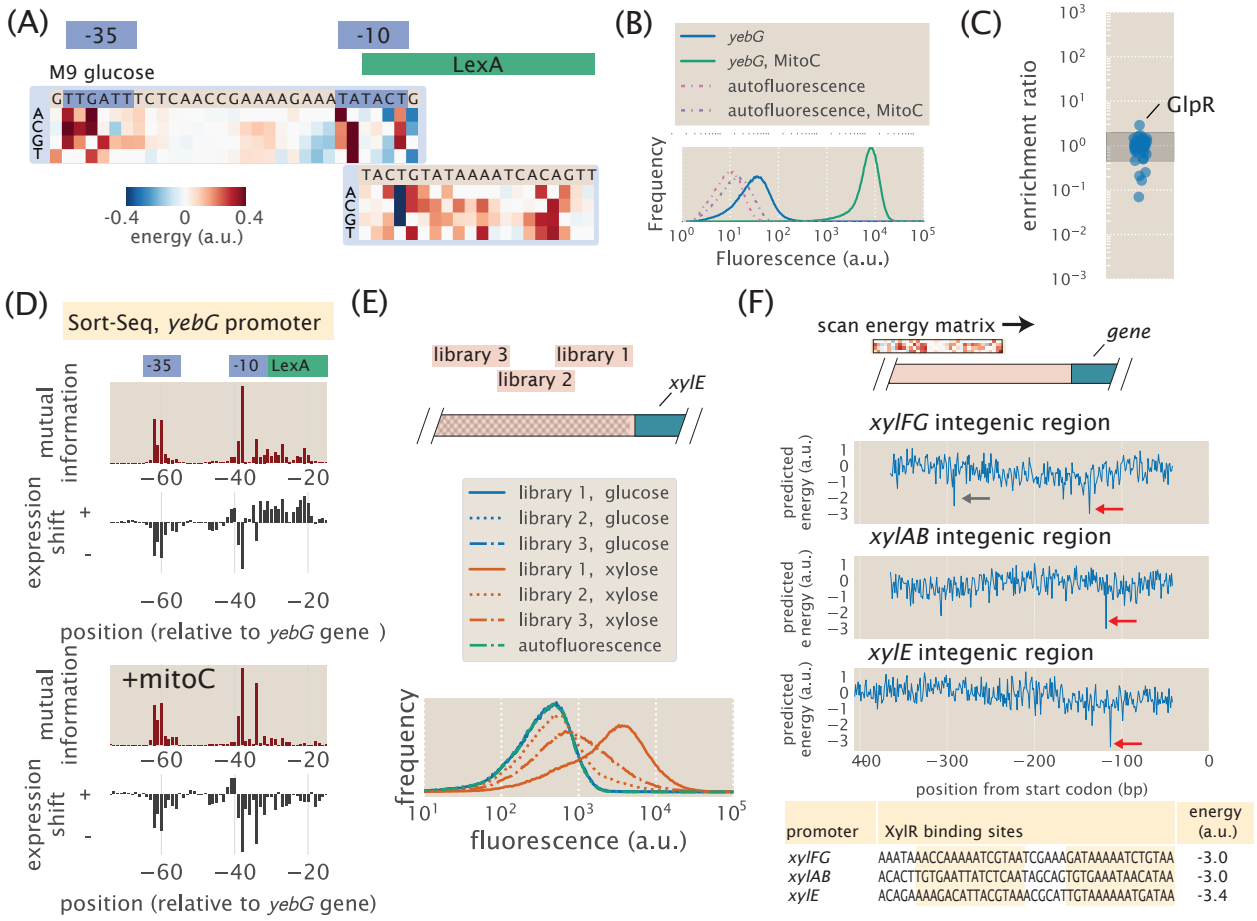

Figure 3.15: **Extended analysis of the *yebG*, *purT*, and *xylE* promoters.** (A) Energy matrices were inferred for the binding sites of LexA and RNAP. Data are from cells grown in M9 minimal media with 0.5% glucose. (B) Fluorescence histograms for a wild-type *yebG* promoter plasmid are shown for cells grown in M9 minimal media with 0.5% glucose, and with or without mitomycin C (1 μg/ml). Mitomycin C induces the SOS response [48] and dramatically increases expression from the *yebG* promoter. Autofluorescence histograms refer to cells that did not contain the GFP promoter plasmid. (C) DNA affinity chromatography performed using the identified repressor site on the *purT* promoter. Cell lysate was produced from cells grown in M9 minimal media with 0.5 % glucose and binding was performed in the presence of adenine (100 μg/ml) to match the growth conditions where repression was observed. (D) Information footprints and expression shift plots are shown for the *yebG* promoter in the presence or absence of mitomycin C (1 μg/ml). Cells were grown in M9 minimal media 0.5% glucose. (E) Fluorescence histograms are shown for the three *xylE* libraries (different mutated regions), with cells grown in M9 minimal media with either 0.5% glucose or 0.5% xylose. While xylose led to differential expression for the different libraries, cells grown in glucose were identical to autofluorescence. (F) The energy matrix associated with two tandem putative binding sites for xylR and CRP (Fig. 6C) was scanned across the intergenic regions of *xylAB*, *xylFG*, and *xylE*. The predicted energy is plotted for each position, and a strong binding site was identified in each promoter (red arrow). For *xylAB*, and *xylFG*, this matched the known binding sites for XylR and CRP on these promoters and their sequences and binding energy predictions are noted below the plots. The promoters of *xylAB* and *xylFG* share the same intergenic regions, but are in opposite coding directions. The reverse complement of the binding site identified in the *xylAB* promoter also showed a strong binding energy prediction (gray arrow in *xylFG* scan).

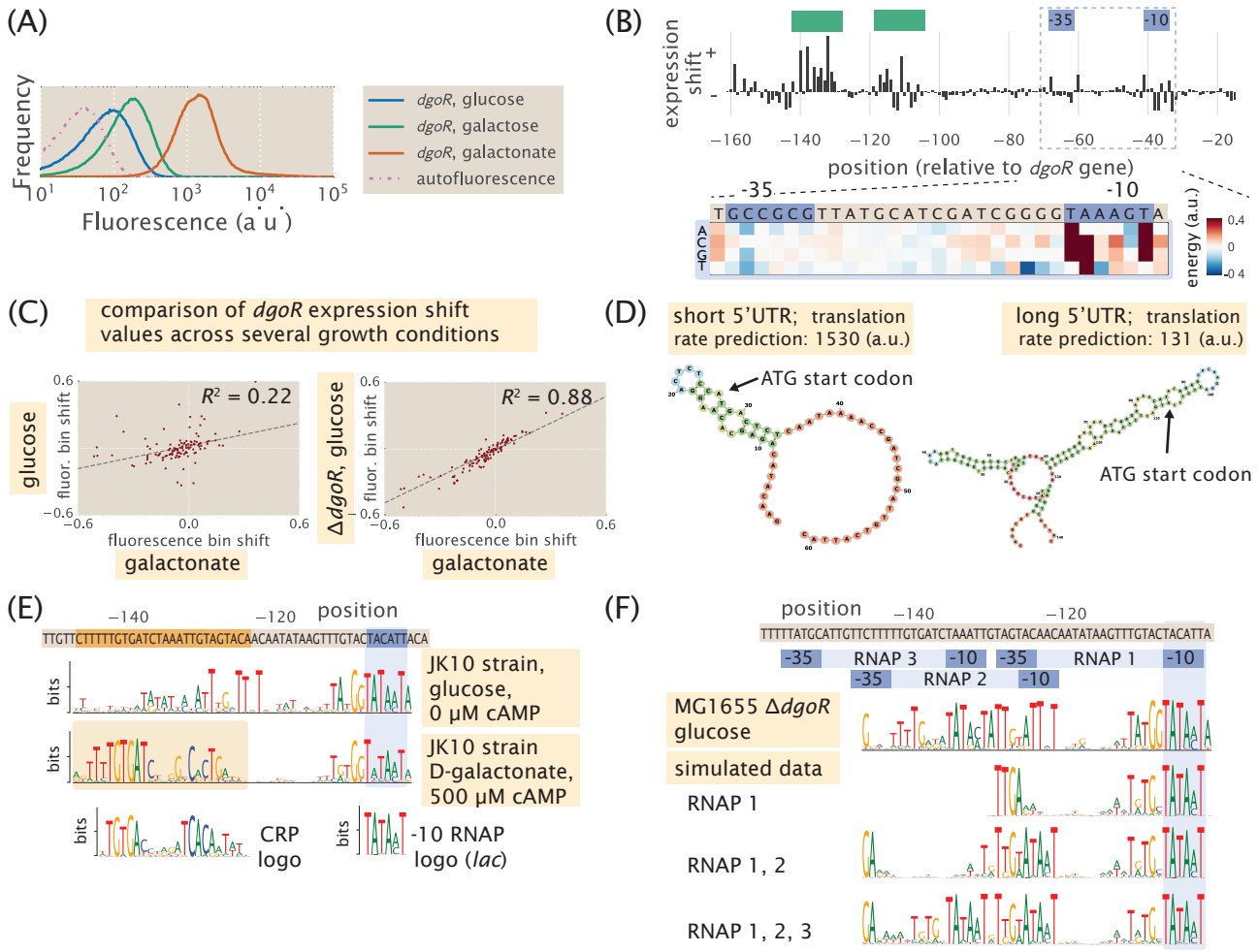

Figure 3.16: **Extended analysis of the *dgoR* promoter.** (A) Flow cytometry histograms of cells containing a wild-type *dgoR* promoter plasmid are shown for cells grown in M9 minimal media with 0.5% glucose, 0.23% galactose, or 0.23% D-galactonate. (B) Identification of an RNAP binding site that appears active when cells are grown in M9 minimal media with 0.5% glucose. The inferred energy matrix exhibits a clear -10 RNAP binding site (consensus sequence is TATAAT) and a poor -35 binding site (CCCCCC). (C) Expression shift values are plotted against each other (glucose vs. D-galactonate, and Δ*dgoR* glucose vs. D-galactonate) for positions -120 bp to -14 bp relative to the *dgoR* coding gene. Note that these are the same values used to generate the plot in Fig. 7A, just plotted against each other for each position. Δ*dgoR* cells appear to have the same regulatory phenotype as cells grown in D-galactonate, with a line of best fit showing much higher correlation between these data sets. (D) Predicted RNA transcript structure based on the two distinct RNAP binding sites. Growth in D-galactonate leads to the long 5' untranslated region and is found to produce strong secondary structure which predicts significantly lower translation rates of the *dgoR* gene than with the short transcript. The ATG start codon is identified. (E) Sequence logos were generated for the most upstream 60bp region containing the putative RNAP and CRP binding sites. Data is from Sort-Seq in strain JK10 (derivative of TK310 [13]) and binding of CRP was induced through addition of 500 μM cAMP. Cells were grown in EZrich MOPS media (Teknova, CA, USA) with D-Galactonate as the carbon source. In comparison to the sequence logos shown in Fig. 7C (growth in D-galactonate), the right side of the CRP binding site is now in better agreement with the logo from the *lac* promoter. (F) Sequence logos are shown for simulated data for the upstream region of the *dgoR* promoter assuming one, two, or three RNAP binding sites. The top sequence logo shows the experimental result for Sort-Seq performed in a Δ*dgoR* genetic background, with cells grown in glucose.

3.12 Supplemental Information: Extended Sort-Seq data analysis details.

Calculation of expression shifts

One of the first ways we analyze the sequence data from our Sort-Seq experiment is to look at the consequence of mutations at each position on the overall fluorescence. Specifically, at each position we calculate the average fluorescence bin of mutated nucleotides and compare this to the average bin for all the sequences in the data set (i.e. expression shift). Since we find that most mutations are deleterious to the binding of transcription factors or RNAP, we can use the change in fluorescence to identify regions associated with binding by repressors or activators and RNAP. This provides an alternative to the information footprints calculated in Kinney et al., 2010. While the information footprints can also be useful, the sign of the expression shifts is useful to determine the type of regulatory protein.

First we calculate the average bin for all the sequences in the data set. We let N_f be the total number of sequences in each bin, where f refers to the bin number (f = 1, 2, 3, and 4, for four bins). The average fluorescence bin is then given by the arithmetic average across all bins,

$$\langle f \rangle = \sum_{f=1}^{4} f \cdot p(f) = \sum_{f=1}^{4} f \cdot \frac{N_f}{\sum_{f=1}^{4} N_f}, \tag{3.12}$$

where $p(f)$ is the fraction of sequences in bin f. Note that the denominator is just the total number of sequences, $N = \sum_{f=1}^{4} N_f$, and that this average will be independent of position.

Next we need to determine the average fluorescence bin of a mutated nucleotide at each position i. Since the number of mutated nucleotides may differ at each position, we define the number of mutated nucleotides in each bin and position as $M_{f,i}$. The subscript 'f, i' is used to identify which bin f and position i are being considered. The average fluorescence bin of a mutated nucleotide can then similarly be found,

$$\langle f_{mut,i} \rangle = \sum_{f=1}^{4} f \cdot p_{mut,i}(f) = \sum_{f=1}^{4} f \cdot \frac{M_{f,i}}{\sum_{f=1}^{4} M_{f,i}}, \tag{3.13}$$

where in this case, $p_{mut,i}(f)$ refers to the fraction of mutated nucleotides in bin f, and at position i.

Finally, we can now calculate the average fluorescence bin shift upon mutation,

which is given by the differences in Equation 3.13 and Equation 3.12,

$$\langle \Delta f_{mut,i} \rangle = \langle f_{mut,i} \rangle - \langle f \rangle == \sum_{f=1}^{4} f \cdot \left(\frac{M_{f,i}}{\sum_{f=1}^{4} M_{f,i}} - \frac{N_f}{\sum_{f=1}^{4} N_f} \right). \tag{3.14}$$

Note that when we plot the fluorescence bin shift for a region where we have multiple data points (i.e. from different mutated, but overlapping regions of the DNA), we plot the average calculated value of $\langle \Delta f_{mut,i} \rangle$ from the different experiments.

We also note that it is possible to re-weight each bin by its mean fluorescence, $\widetilde{f}$ (i.e. instead of $f = 1, 2, 3, 4$, use the average fluorescence shift in arbitrary fluorescence units). Here we replace f with $\widetilde{f}$ in Equation 3.14. For example, under situations where different sort conditions were used across experiments, this re-normalization should allow better comparison of values across experiments. The fluorescence values for $\widetilde{f}$ can be determined by regrowing the sorted cells and measuring the mean fluorescence of each sorted cell population.

Calculation of information footprints

Another way that we analyze the data from our Sort-Seq experiments is to calculate an information footprint [13]. This allows us to identify whether there are any positions along the mutagenesis window that are informative in relating sequence S and fluorescence bin f. Said differently, an informative region would be one that if given some knowledge about the sequence, we should be able to predict which fluorescence bin the promoter sequence might be found in. The mathematical way of implementing this intuition is to use the quantity known as the mutual information.

We can calculate the mutual information between sequence and fluorescence bin, $I(b_j, f)$, at each position i along the mutagenesis window by calculating the fraction of each nucleotide b_j (= A, C, G, T) found within each bin f. This allows us to estimate the joint probability distribution $p_i(b_j, f)$ at each position i. For example, $p_{10}(A, 2)$ would denote the probability that we observe an A in the second fluorescence bin at position i=10 along our promoter. The mutual information at each position is then defined by

$$I_i(b_j, f) = \sum_{b_j=A}^{T} \sum_{f=1}^{N_f} p_i(b_j, f) \log \left(\frac{p_i(b_j, f)}{p_i(b_j) p_i(f)} \right), \tag{3.15}$$

where we have summed over all nucleotides and the N_f fluorescent bins that the sequences were found in. There is also a finite sample correction that can be applied,

[99], since Equation 3.15 tends to overestimate the true mutual information. This is given by

$$I_i(b_j, f) = \sum_{b_j=A}^{T} \sum_{f=1}^{N_f} p_i(b_j, f) \log\left(\frac{p_i(b_j, f)}{p_i(b_j)p_i(f)}\right) - \frac{(n_{b_j} - 1) \cdot (n_f - 1) \cdot log_2 e}{2 \cdot N} + O(N^{-2}), \tag{3.16}$$

where $n_{b_j} = 4$ is the number of nucleotides, and n_f is the number of bins that cells have been sorted into.

Inference of energy matrix models with Sort-Seq data.

In order to predict the influence of DNA sequence on binding by regulatory proteins, we use the Sort-Seq data to generate quantitative models of the sequence-dependent binding energy. Through a relationship between likelihood and mutual information, Kinney *et al.* [13, 100] showed that in the large data limit it is possible to infer biophysical parameters such as the binding energies that relate the interaction between proteins and DNA sequence. In this section we describe in detail the approach used to infer energy matrices from our Sort-Seq data using Markov Chain Monte Carlo (MCMC). A full discussion of MCMC is beyond the scope of this work, but we point the interested reader to further details regarding inference using mutual information in work from Kinney *et al.* [13, 52, 98]. We also stress that while we make extensive use of linear energy matrix models, the inference procedure is in no way limited to such models and can be extended to allow, for example, epistatic effects through the addition of other parameters. The simple linear models, however, provide us with a useful starting point to gain insight and describe the protein-DNA interaction.

We begin with a summary of the procedure used to infer an energy matrix model using MCMC, and use the RNAP binding site of the *relB* promoter as an example. The inference was performed using the MPAthic software [25]. A general schematic of the procedure is shown in Figure 3.17. More specific details are then discussed in the following subsections. First we must initialize a 4xL set of energy parameters, $\Theta = \{\theta_{i,j}\}$, for a binding site of length L and four base pairs (see Figure 3.17, part 1). We begin by randomly selecting parameter values for our energy matrix with which to initialize the MCMC. Here we select values from a normal distribution centered at zero with variance equal to 1, although this choice does not appear to be too critical, and rather just provides us with a starting point for our MCMC chain. Using this energy matrix we then estimate the mutual information between

the binned sequences and the associated set of energy model predictions. As shown in Figure 3.17, part 2, initially the energy matrix will be of little value in describing the observed sequence data since it was randomly chosen. This is shown by the almost uniform joint probability distribution and low mutual information in Figure 3.17A, and Figure 3.17B.

We now begin the MCMC by perturbing the energy matrix parameters using the Metropolis-Hastings algorithm with the PyMC package in Python [101] (within the MPAthic software [25]). After each step of the chain, we re-calculate the mutual information between the data and new model predictions, which allows us to calculate how well this new set of energy matrix parameters describe the data. Dependent on whether the new energy matrix parameters lead to an improvement in mutual information, these new parameters are either retained or discarded and the process is repeated (again, according to the Metropolis-Hastings algorithm [101]). We also renormalize the matrix entries to constrain certain gauge freedoms after each iteration.

After a sufficient number of steps, and assuming that a model exists that can describe the Sort-Seq data, we will arrive at a model whose joint probability distribution between model predictions and binned sequences show a clear correlation. This is shown by the joint probability distribution in Figure 3.17C, as well as the plateau in the mutual information trace in Figure 3.17A, since changes to the energy matrix parameters are unable to increase the mutual information any further. In this first portion of MCMC we have performed many samplings to reach a high probability region where the energy matrix will be more representative of the distribution we are sampling from. This first step is usually referred to as the 'burn-in' period [101] and allows us to begin sampling from the distribution, $p(\Theta|data)$, that describes the distribution of energy matrix model parameters.

Finally, now that we are sampling from the desired distribution, we can estimate energy matrix parameters just by sampling this distribution many times. This brings us to part 3 of Figure 3.17. While the mutual information no longer shows a substantial change, the parameters of the energy matrix are continuing to be perturbed following the Metropolis-Hastings algorithm, and according to the distribution $p(\Theta|data)$. We can now estimate each entry in the energy matrix by taking the arithmetic mean of the matrix parameters across all the sampling steps. This is shown by a set of contour plots and marginalized distributions for the binding energy parameters from column five of the RNAP energy matrix (Figure 3.17D). To ensure that multiple

energy minima were not present in this energy landscape, we repeated the inference procedure 20 times and used the average across all appropriate MCMC chains to estimate the energy matrix parameters. The calculated mutual information will be indifferent the particular sign of the energy matrix and adjust the energy matrices such that the wild-type sequence has a negative predicted binding energy and check that energy predictions from the energy matrices from each MCMC are correlated (keeping energy matrices that provide a Pearson correlation coefficient of 0.85 or greater across model predictions). Note that for inference of parameters using thermodynamic models, separate from these energy weight matrices, we did find the presence of multiple minima and apply a parallel tempering MCMC procedure to properly sample these distributions.

Using the schematic in Figure 3.17 as our guide, the sub-sections that follow expand on the details introduced here to perform this inference procedure. In particular, we begin by describing the linear energy matrix model. We then outline the Bayesian approach taken to formally write the posterior distribution, $p(\Theta|data)$, that provides us with a relationship between the energy matrix parameters and observed sequence data. When sampling this distribution we need to estimate mutual information at each iteration of the MCMC sampling procedure, and describe how to calculate later in this section.

Linear energy matrix models are used to describe DNA-protein interaction.

We begin by outlining the linear energy matrix model shown in Figure 3.17A that describes the binding interaction between the DNA and a DNA-binding protein. We treat each base pair position j along a binding site as contributing a certain amount to the binding energy, where the total binding energy is then the sum of the contributions from all base pairs. Mathematically the energy matrix model is described by a 4xL matrix, Θ, consisting of energy parameters $\{\theta_{ij}\}$. Here each column j of matrix parameters will represent the energies for each nucleotide i = A, C, G, or T (= $1, 2, 3$, or 4) associated with position j of the binding site. For example, $\theta_{2,3}$ represents the energy parameter for nucleotide C at position 3. To make our computation of binding energies more convenient, we also represent our DNA sequence as another matrix, S, having identical dimensions, 4xL. This matrix consists of parameters $\{s_{ij}\}$, where the ij^{th} entry again corresponds to the the nucleotide identity i and sequence position j. Each parameter will have a value of 1 if it corresponds to the sequence's nucleotide identity at position j, and a value

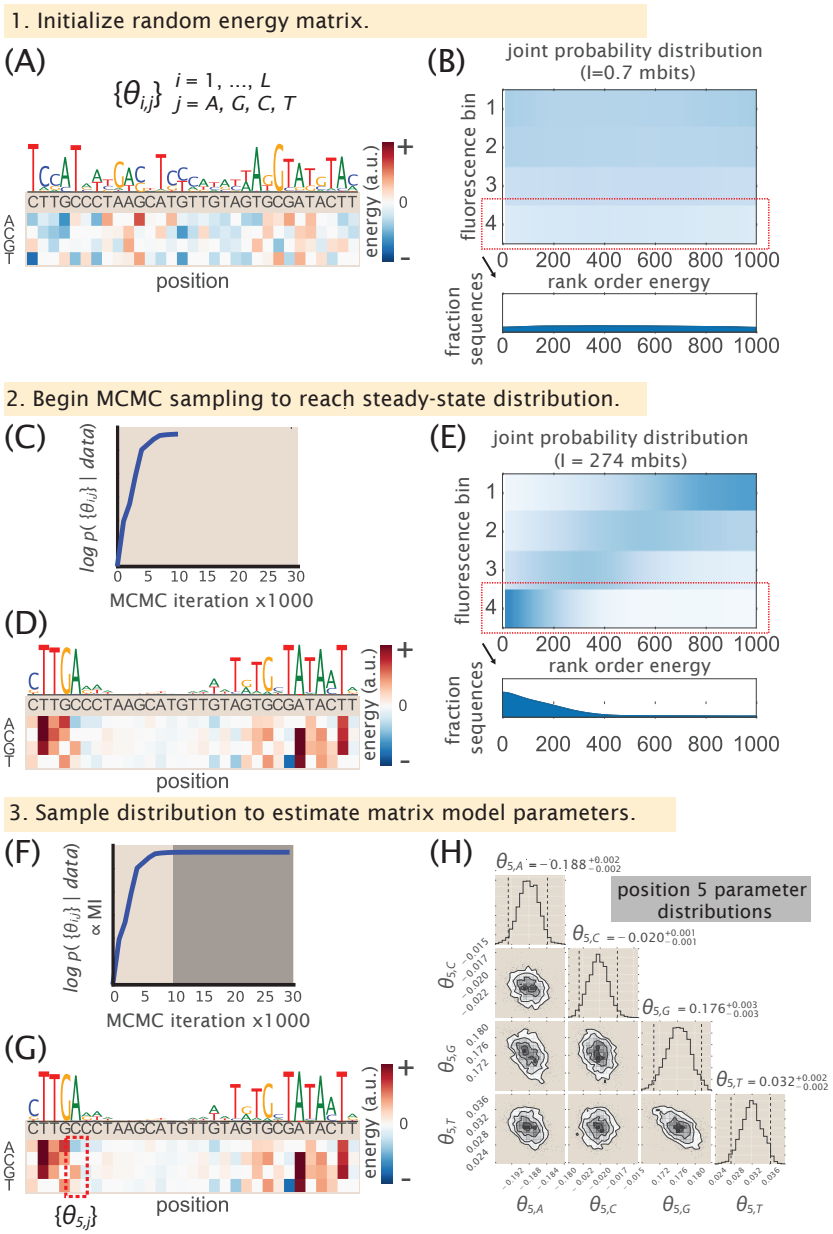

Figure 3.17: **Schematic of the inference procedure used to determine energy matrices from Sort-Seq data using Markov Chain Monte Carlo.** 1. To begin the inference of a set of 4xL model parameters, $\{\theta_{ij}\}$, are chosen from a normal distribution. (A) Example set of parameters used to initialize the MCMC sampling. Matrix entries are first normalized such that energy predictions have mean of zero and standard deviation of one. For plotting energy matrices, each column has been shifted such that the wild-type sequence has zero energy. The associated sequence logo is shown above the energy matrix. (B) Estimated joint probability distribution between fluorescence bin and rank order energy predictions using the energy matrix in (A), using all sequences in the *rel* promoter data set. The bottom plot shows, the histogram of rank ordered predictions of only bin four, corresponding to the red boxed region, which is nearly uniform due to the randomly chosen matrix entries used to predict energies from each sequence. Since the matrix parameters were randomly chosen, the nearly uniform distribution results in low mutual information (0.7 mbits, where 1 mbit = 10^{-3} bits) between fluorescence bin and rank order energy predictions. *(Caption continued on next page)*

Figure 3.17: *(continued from previous page)* 2. MCMC sampling of the energy matrix model is performed using the Sort-Seq data associated with the *rel* RNAP binding site. (C) The log posterior, Equation 3.20, is plotted for the first 1000 iterations and corresponds to the "burn-in" period. The log posterior is proportional to the mutual information between fluorescent bin and rank order energy predictions. During each sampling iteration, the parameters will be retained or discarded with some probability given by the the Metropolis-Hasting algorithm [101]. (D) The energy matrix and sequence logos are shown using the set of parameters at the 1000^{th} iteration. (E) Estimated joint probability distribution between fluorescence bin and rank order energy predictions using the energy matrix in (D). The energy matrix provides energy predictions for each sequence that clearly distributes across the sorted bins and results in much higher mutual information (274 mbits). 3. Finally, matrix parameters are estimated by continuing to sample the posterior distribution many more times and determined from a weighted average of these samples. (F) The log posterior is plotted for the entire set of MCMC iterations. The sampled model parameters during the shaded region are used to estimation each matrix entry. (G) The mean energy matrix entries from these samples are plotted. (H) Contour plots and marginalized distributions summarize the sampled values for each of the four parameters at position five of the RNAP energy matrix. Note that entries in (G) have been shifted such that the wild-type nucleotide has zero energy.

of 0 otherwise. For example, for a sequence with a C at position $j = 4$, the entry $s_{2,4} = 1$ and $s_{i=1,3,4,j=4} = 0$. The binding energy, E, of any sequence, S, will then be given by

$$E = \sum_{i=A}^{T} \sum_{j=1}^{L} \theta_{ij} \cdot s_{ij}. \tag{3.17}$$

One aspect we have not considered thus far is the scale of the energy parameter. When considering binding between between DNA and a DNA-binding protein, a statistical mechanical approach would suggest that the probability of such an event occurring will be given by the Boltzmann factor, $e^{-\varepsilon_s/(k_B T)}$ [51]. Here ε_s is the binding energy that describes this interaction in absolute energy units (e.g. units of $k_B T$; 1 kcal/mol = 1.62 $k_B T$ at 37°C), k_B is the Boltzmann constant, and T is temperature. In relation to the binding energy, E, described by our Equation 3.17 above, $\varepsilon_s = A \cdot E + B$, where the constant A scales the energy matrix into absolute energy units, while B provides an additive shift that depends on the choice of reference energy. Here, the matrix entries that are used to calculate E are 'gauge fixed' such that the mean energy in each column is set to zero and the matrix norm (or inner product) has a value of 1. Note however that when plotting each energy

matrix we find it useful to shift the energy in each column such that the wild-type sequence has zero energy.

When fitting the data to a model of the form $e^{-\varepsilon_s/(k_B T)}$, the fitting procedure is unable to determine the scale factors A and B noted above. For example, in most instances we report energy values in arbitrary units. This is consequence of the fitting procedure, where in the absence of a specific thermodynamic model, there remain some scale parameters that cannot be determined [13]. This parameter insensitivity has been termed 'diffeomorphic modes' and is discussed at length in other work [52]. One especially interesting aspect of this is that when considering biophysical models of regulation, diffeomorphic modes often disappear and make it possible to infer parameters that were not accessible by fitting simpler models. For the cases of repression by PurR at the *purT* promoter, or activation by CRP at the *dgoR* promoter, this allowed us to estimate binding energy in absolute energy. We discuss this further later in this section, and in Chapter 4 we consider energy matrix scaling in more detail.

Probability distribution relating energy matrix model parameters to the Sort-Seq data.

Given our FACS-sorted sequence data, we want to find the set of energy matrix parameters that best describe the distribution of sequences across our fluorescence bins (i.e. parameters that provide binding energy predictions that describe the data as shown in Figure 3.17C). To perform this inference we take a Bayesian approach in our analysis, and as mentioned earlier, rely on MCMC to sample from the complex distribution relating our energy matrix parameters to the sequence data. While a full discussion of Bayesian analysis is outside the scope of this section, the book, Data Analysis by Sivia and Skilling [102], and online material available from the Caltech course, *BE/Bi 103: Data analysis in the biological sciences*, taught by Justin Bois (`http://bois.caltech.edu/teaching.html`), are excellent resources.

Formally, we want to find the set of energy matrix parameters that maximize the probability distribution of our energy predictions (through our energy matrix model) given our Sort-Seq sequence data, $p(E|\{S, f\})$, where $\{S, f\}$ refers to our array of N sequences S and the bin f where they were found (referred to as the 'data' in the initial summary of the inference procedure). x_S is the binding energy as defined in

Equation 3.17. From Bayes' theorem, we can re-write this distribution as

$$p(E|\{S,f\}) = \frac{p(\{S,f\}|E)p(E)}{p(\{S,f\})} \propto p(\{S,f\}|E)p(E), \tag{3.18}$$

where the term $p(\{S,f\}|E)$ is called the likelihood, and $p(E)$ is known as the prior and encompasses our prior knowledge on the energy matrix parameters. The denominator $p(\{S,f\})$ is known as the marginalized likelihood and acts as a normalization factor, but is unimportant for our inference.

To proceed we follow the approach of Kinney *et al.* [13, 100]. We assume a uniform prior over the energy matrix model parameters. In addition, we also assume our sequence measurements are independent. The second assumption allows us to write $p(\{S,f\}|E)$ as the product of probabilities across all sequences contained within our data set, $p(\{S,f\}|E) = \prod_{s=1}^{N} p((S_i,f_i)|E)$. This is also referred to as the error model since by relating the binned sequence data to binding energy, it must also encompass the additional noise sources from our experiment that actually led to our array of sequence data. Noise sources that might influence this include the sensitivity of the FACS GFP measurements, and the rate of mis-sorting events. Expression variability due to stochastic gene expression, differences in cell size, and plasmid copy number fluctuations are also likely to contribute. However, since these are not known exactly, Kinney *et al.* computed the likelihood by averaging over an ensemble of all possible error models. Using a uniform prior over the possible error models they found,

$$p(\{S,f\}|E) = \left\langle \prod_{s=1}^{N} p((S_i,f_i)|E) \right\rangle_{\text{all possible } p(S_i,f_i|E)} = C \cdot 2^{N\cdot(I(f,E)+\Delta)}, \tag{3.19}$$

where N is the total number of sequences considered, $I(f,E)$ is the mutual information between the observed fluorescence bins and binding energies predicted by the energy matrix for all the sequences, and C is a constant of integration that will be unimportant to us. Here, Δ is a small correction that goes to zero as N goes to infinity [100]. Inserting Equation 3.19 into Equation 3.18, we can write

$$p(E|\{S,f\}) \propto 2^{N\cdot I(f,E)}. \tag{3.20}$$

Here we have assumed that N is sufficiently large so that the prior (which does not scale with N), as well as the Δ term in Equation 3.19 can be ignored. To reiterate

in reference to our MCMC procedure (shown in Figure 3.17), this is the probability distribution that we are sampling from to find the set of energy matrix parameters that describe our sorted sequence data set. The mutual information values shown in the plots of Figure 3.17C, F (mutual information traces in part 2 and 3) are reflected by our choice of energy matrix parameters. MCMC enables us to sample from the distribution and essentially find the set of matrix parameters that maximize this mutual information. In the next section we continue by describing how we estimate mutual information.

Estimating mutual information using the energy model predictions.

In the last section we found that the energy matrix parameters should be related to the data through Equation 3.20. By performing many samples from this distribution using MCMC, it is possible to estimate the most probable energy matrix parameters, $\theta_{i,j}$, that make up our energy matrix. Here we consider how to estimate the mutual information term in Equation 3.20 needed for our calculation. While a non-trivial problem in general, the following approach appears to work well in practice. In this case the fluorescence bins, f, are discrete variables while our binding energies, E, are continuous, with the mutual information given by

$$I(f,E) = \int_{E=-\infty}^{E=\infty} dE \sum_{f} p(f,E) \log_2 \frac{p(f,E)}{p(E) \cdot p(f)}. \tag{3.21}$$

In our sequence data set, we can easily estimate $p(f)$ by counting the number of sequences in each fluorescence bin. However, we do not have direct access to the probability distribution $p(E)$ *a priori*.

To proceed, we further bin our N sequences into 1000 bins, by rank ordering them by their associated binding energy predictions (using the energy matrix of the current MCMC step). This provides us with an estimate of the probability distribution in binding energy across our sequences. Specifically, this is shown for fluorescence bin 4 in Figure 3.17B and E. While this is not a direct estimate of $p(E)$, we invoke the fact that the mutual information will be invariant under monotonic transformations ($I(f,E) = I(f,z_s)$) [13]. Hence, instead of calculating $I(f,E)$, we instead calculate $I(f,z_s)$, where z_s is instead the ranked ordering of the N sequences.

In order to calculate the mutual information we now construct a 2-d histogram (joint distribution) by binning the rank ordered energy predictions into $z_s = 1$ to 1000 bins across each of the different fluorescence bins. We define this by the frequency

matrix $F(f, z_s)$, and from our finite data set, use kernel density estimation with a kernel width equal to 4% to estimate the joint distribution. This is what is plotted in Figure 3.17B, and E, where the mutual information is then calculated as

$$I(f, z_s)_{smooth} = \sum_{z_s=1}^{1000} \sum_{f} F(f, z_s) \log_2 \frac{F(f, z_s)}{F(z_s) \cdot F(f)}. \tag{3.22}$$

Inference of thermodynamic model parameters using parallel tempering Markov chain Monte Carlo (MCMC).

So far, we have applied MCMC using an error-model-averaged likelihood to infer the parameters of an energy matrix. One limit initially observed by Kinney *et al.* [13] was an inability of the fitting procedure to constrain certain parameters (due to free diffeomorphic modes, noted earlier). Interestingly however, it was found that certain diffeomorphic modes often disappear when fitting the Sort-Seq data to non-linear models. For a thorough discussion of diffeomorphic modes refer to the work of Kinney *et al.* [103]. We applied this strategy in several of our data sets from the *purT*, *dgoR*, and *xylE*, where specific thermodynamic models appeared appropriate. Here we briefly outline the models used and the main results from our MCMC analysis.

We begin with the *purT* promoter. Here we identified an RNAP binding site that is repressed by PurR, which binds between the -10 and -35 RNAP sites. Given the presence of only these two binding sites, we modeled the promoter as having a simple repression architecture [51]. Some additional complexity arises due to the presence of other PurR binding sites on the genome, and the allosteric dependence of a purine metabolite for co-repression. Following the approach of Weinert *et al.* [90], this can be quantitatively described by

$$P_{bound} = \frac{\lambda_p e^{-\beta \varepsilon_p}}{1 + \lambda_p e^{-\beta \varepsilon_p} + \lambda_r e^{-\beta \varepsilon_r}}. \tag{3.23}$$

Here λ_p and λ_r represent the fugacity, which describes the relative availability of RNAP and PurR, respectively, to bind their binding sites. These parameters depend on the concentration of each protein (through their chemical potentials), and for PurR, will also depend on its allosteric state. ε_p and ε_r represent the binding energies of RNAP and PurR to their binding sites, respectively.

We can also describe each binding energy through the gauge-fixed energy matrix prediction, which is multiplied by a scale factor and additive shift (e.g. $\varepsilon_r = A_r \cdot x_r + B_r$, where A_r is the scale factor, x_r is the energy matrix prediction, and B_r is the additive shift). To being fitting to the model described by Equation 3.23, we first inferred the energy matrices for RNAP and PurR following the MCMC procedure noted above. We then performed a second MCMC to fit the remaining thermodynamic parameters. In this second MCMC we sampled using error-model-averaged likelihood against the posterior $p(P_{bound}|\{S, f\})$. This allowed us to infer the following parameters: $A_r = -11.55^{+0.2}_{-0.5} k_B T$, $\lambda_r e^{-\beta B_r} = e^{0.64^{+0.1}_{-0.3}}$, and $A_p = 2.4^{+0.4}_{-0.1} k_B T$, where A_p is the RNAP scale factor. Here the error bars represent the median of their respective distributions, where the superscripts and subscripts represent the upper and lower bounds of the 95$^{\text{th}}$ percentile of the parameter value distributions. Note that in this second MCMC, we performed parallel tempering MCMC (using the PTSampler in package emcee, [104]) to better sample the posterior distributions of our thermodynamic parameters (see supplemental material of Kinney *et al*, 2010).

Next we consider the *dgoR* promoter. While we found the promoter to be quite complex, here we use data from the JK10 strain (see Supplemental Section 3.11) where activation by CRP appeared to dominate transcription. Here we apply the model used by Kinney *et al.* [13], which consists of a binding site for RNAP and CRP, but also includes an interaction energy between these two proteins. Again using fugacity terms to describe the availability of each protein, this will be given by

$$P_{bound} = \frac{\lambda_p e^{-\beta\varepsilon_p} + \lambda_a \cdot \lambda_p e^{-\beta(\varepsilon_p+\varepsilon_a+\varepsilon_i)}}{1 + \lambda_p e^{-\beta\varepsilon_p} + \lambda_a e^{-\beta\varepsilon_a} + \lambda_a \cdot \lambda_p e^{-\beta(\varepsilon_p+\varepsilon_a+\varepsilon_i)}}. \tag{3.24}$$

In this architecture we have the fugacity λ_a for the activator CRP and its binding energy to the binding site, ε_a. In addition, there is an additional energy term ε_i that describes the interaction between RNAP and CRP. Again, we can write $\varepsilon_p = A_p \cdot x_p + B_p$. We can also write the CRP binding energy as $\varepsilon_a = A_a \cdot x_a + B_a$, where similarly, A_a is the scale factor, x_a is the gauge-fixed energy prediction, and B_a is an additive shift. Using parallel tempering MCMC to sample $p(P_{bound}|\{S, f\})$, we obtained the following values: $\varepsilon_i = -7.3^{+1.9}_{-1.4} k_B T$, $A_a = -13.6^{+2.6}_{-2.2} k_B T$, $\lambda_a e^{-\beta B_a} = e^{-1.89^{+0.4}_{-0.6}}$, and $A_p = -12.7^{+3.4}_{-2.8} k_B T$. As with the *purT* case above, the error bars represent the median of their respective distributions, where the superscripts and subscripts represent the upper and lower bounds of the 95$^{\text{th}}$ percentile of the parameter value distributions.

Lastly we consider the *xylE* promoter. This promoter contains two XylR sites which are likely bound as a dimer [53]. There is also a CRP site directly upstream of the xylR sites. The binding signature of CRP is only observed for the right half of the binding site, implying the left half of the protein does not make as significant DNA contact. Since CRP still has a powerful impact on gene expression, it suggests that there is a cooperative interaction between xylR and the weak CRP site. The short distance between the xylR sites and the RNAP also suggests that there is a direct interaction between the xylR sites and the RNAP. In addition, there is also a spacing between the RNAP polymerase and the CRP site of 35 bp (approximately three helical turns of the DNA). For this spacer length in the *lac* promoter there is a expected to be a significant interaction energy even in the absence of XylR [105, 106]. A thermodynamic model of RNAP polymerase binding probability for this architecture will be

$$P_{bound} = \frac{f(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i})}{g(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i})}, \tag{3.25}$$

where

$$\begin{aligned} f(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i}) &= \lambda_p e^{-\beta\varepsilon_p} + \lambda_p\lambda_x e^{-\beta(\varepsilon_p+\varepsilon_x+\varepsilon_{x_i})} \\ &\quad +\lambda_p\lambda_c e^{-\beta(\varepsilon_p+\varepsilon_c+\varepsilon_{c_i})} + \lambda_p\lambda_c\lambda_x e^{-\beta(\varepsilon_p+\varepsilon_x+\varepsilon_c+\varepsilon_{c_i}+\varepsilon_{x_i}+\varepsilon_{cx_i})} \end{aligned} \tag{3.26}$$

$$\begin{aligned} g(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i}) &= 1 + \lambda_x e^{-\beta\varepsilon_x} + \lambda_c e^{-\beta\varepsilon_c} + \lambda_x\lambda_c e^{-\beta(\varepsilon_x+\varepsilon_c+\varepsilon_{cx_i})} \\ &\quad +\lambda_p e^{-\beta\varepsilon_p} + \lambda_p\lambda_x e^{-\beta(\varepsilon_p+\varepsilon_x+\varepsilon_{x_i})} \\ &\quad +\lambda_p\lambda_c e^{-\beta(\varepsilon_p+\varepsilon_c+\varepsilon_{c_i})} \\ &\quad +\lambda_p\lambda_c\lambda_x e^{-\beta(\varepsilon_p+\varepsilon_x+\varepsilon_c+\varepsilon_{c_i}+\varepsilon_{x_i}+\varepsilon_{cx_i})}. \end{aligned} \tag{3.27}$$

Here, the λ_x and ε_x terms mark the fugacity and binding energy of XylR respectively. The λ_c and ε_c represent the fugacity and binding energy of CRP, and λ_p and ε_p do the same for RNAP. The terms ε_{x_i}, ε_{c_i}, and ε_{cx_i} are interaction terms between XylR and RNAP, CRP and RNAP, and CRP and XylR, respectively.

Due to the position of the library windows (with a 60 bp window containing the two XylR binding sites, but only partial binding sites for CRP and RNAP), we were unable to fit this model to the data. The fitting procedure requires sequences with mutations throughout the multiple binding sites and further experimentation will be needed to fit and characterize the proposed model further.

3.13 Supplemental Information: Extended experimental details

In this section we provide additional details to describe the specifics of the work flow. In general, an experiment is begun by constructing the mutated promoter libraries for Sort-Seq. Next transform libraries into cells and use FACS to sort by fluorescence. Using putative regulatory sequences identified by Sort-seq, we perform DNA affinity chromatography and mass spectrometry, which is necessary to identify the transcription factors that bind to these putative binding sites.

E. coli strain construction

Here we describe the approach used to generate these deletion strains. Briefly, an overnight culture of MG1655 containing the plasmid pSIM6 was diluted 1:100 in 50 ml LB media and grown to an OD600 of ≈ 0.4 at 30°C. The culture was immediately placed in a water bath shaker at 43°C for 15 minutes and then cooled in an ice bath for 10 minutes. Cells were then spun down for 10 minutes (4,000 g, 4°C) and resuspended on ice in 50 ml of chilled water. This was repeated three times before resuspending in 200 μL of chilled water to generate competent cells. Homologous primer extension sequences for the appropriate gene were obtained from Baba *et al.* [67] and used to generate linear DNA containing a kanamycin resistance gene insert by PCR, which contained homology for the region on the chromosome to be deleted [78]. Electroporation of the competent cells was performed using 1 μL purified PCR product (about 100 ng DNA), mixed with 50 μL cells. Cells were immediately resuspended in 750 μL SOC media and placed on a shaker at 30°C for outgrowth, for 90-120 minutes. Cells were then plated on an LB-agar plate containing kanamycin (30 μg/ml) and grown overnight at 30°C. The deletions were confirmed by both colony PCR and sequencing. After confirmation, the deletion was transferred to a clean MG1655 strain through P1 transduction and selection on kanamycin. In the case of the lysine auxotrophic strain, we also confirmed deletion of *lysA* by checking that the cells were unable to grow in M9 minimal media unless lysine was supplemented (40 μg/ml).

To generate strains with different LacI tetramer copy numbers per cell (associated with data in Figure 3.12C), the LacI constructs from Garcia *et al.* [34] were P1 transduced into the $\Delta lacIZYA$ strain (integrated at the *ybcN* locus).

Sort-Seq library construction

Mutagenized single-stranded oligonucleotide pools were purchased from Integrated DNA Technologies (Coralville, IA), with a target mutation rate of 9%. Note that in

the case of the *lacZ* promoter, the library is identical to that used in the experiments of Razo-Mejia *et al.* [107], and had a mutation rate of approximately 3%.

Note that to assemble PCR amplified library inserts with the plasmid backbone, we used Gibson assembly [108] (New England Biolabs, MA, USA). Otherwise, we follow the approach of Kinney *et al.* and amplify the backbone using a template plasmid containing the toxic gene *ccdB* (located where the library was to be inserted). This helped ensure that no template plasmid was propagated into the final plasmid library (see methods in reference [13] for more detail).

For each library construction, 40 ng of insert and 50 ng of backbone were combined in a 20 μL Gibson assembly reaction. To achieve high transformation efficiency, reaction buffer components from the Gibson Assembly reaction were removed by drop dialysis and cells were transformed by electroporation of freshly prepared cells. Following an initial outgrowth in 1 mL of SOC media, cells were diluted into 50 mL of LB media and grown overnight under kanamycin selection. Transformation typically yielded $10^6 - 10^7$ colonies as assessed by plating 100 μL of cells diluted 1:10^4 onto an LB plate containing kanamycin.

Sort-Seq experiments

Cells were grown to saturation in LB and then diluted 1:10,000 into the appropriate growth media for the promoter under consideration. For cells grown in 0.23% D-galactonate in M9 minimal media, D-galactonate appeared to form precipitates, but cells otherwise appeared to grow normally. Upon reaching an OD600 of about 0.3, the cells were washed two times with chilled PBS by spinning down the cells at 4000 rpm for 10 minutes at 4°C. After washing with PBS, they were then diluted twofold with PBS to an OD of 0.1-0.15. This diluted cell solution was then passed through a 40 μm cell strainer to eliminate large clumps of cells.

A Beckman Coulter MoFlo XDP cell sorter was used for all Sort-Seq experiments. Prior to sorting, we would obtain fluorescence histograms using between 200,000 and 500,000 cell events per culture. These histograms were used to set the four binning gates, which each covered $\sim$ 15% of the histogram. During sorting of each library, 500,000 cells were collected into each of the four bins. Finally, sorted cells were re-grown overnight in 10 ml of LB media, under kanamycin selection.

Sort-Seq sequencing

The plasmid from cells in each bin were miniprepped following overnight growth (Qiagen, Germany). PCR was used to amplify the mutated region from each plasmid for Illumina sequencing, adding Illumina adapter sequences and custom barcode sequences. Sequencing was performed by either the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech (HiSeq 2500) or NGX Bio (NextSeq sequencer; San Fransisco, CA). Single-end 100bp or paired-end 150bp flow cells were used, with a target read count of about 500,000 sequences per library bin. Joining of paired-end reads was performed with the FLASH tool [109]. For quality filtering, we collected sequences whose barcodes had a PHRED score greater than 20 at each position. Some libraries also contained non-mutagenized regions, and sequences that did not contain the expected sequence were excluded from our analysis. The total number of useful reads available to produce expression shift plots, energy weight matrices, and sequence logos from each Sort-Seq experiment generally ranged between 300,000 to 2,000,000 reads. Energy matrices were inferred using Bayesian parameter estimation with an error-model-averaged likelihood as previously described [13, 52], using the MPAthic software [25]. A more detailed description of the data analysis procedures is available in Supplemental Section 3.12.

DNA affinity chromatography and mass spectrometry

Here we provide additional details on SILAC incorporation, preparation of DNA-tethered magnetic beads, and the LC-MS/MS method.

Lysate preparation and SILAC incorporation

SILAC labeling [27, 28, 30] was implemented by growing cells in either the stable isotopic form of lysine ($^{13}C_6H_{14}{}^{15}N_2O_2$), referred to as the heavy label, or natural lysine, referred to as the light label. By differentially labeling cell lysates we were able to simultaneously quantify the abundance of protein between two DNA affinity purification samples (i.e. one using a target binding site sequence and another as a reference control). This allows us to identify whether any protein shows a preference for the target binding site sequence. Cell lysates were prepared using MG1655 $\Delta lysA$ cells. For each heavy and light labelled cells, 500 ml M9 minimal media was inoculated 1:5,000 with an overnight LB culture of $\Delta lysA$ cells, and grown to an OD600 of ≈ 0.6 (supplemented with the appropriate lysine; 40 μg/ml). Cultures were pelleted, and lysed using a Cell Disruptor (CF Range, Constant Systems Ltd.,

UK) and concentrated to ~150 mg/ml using Amicon Ultra-15 centrifugation units (3kDa MWCO, Millipore).

To generate each lysate an overnight starter culture of cells was grown in LB media supplemented with kanamycin (30 μg/ml). An aliquot was washed twice in M9 minimal media and resuspended to an OD600 of ≈1.0. For both heavy and light labeling, 500 ml M9 minimal media was then inoculated at 1:5,000 and grown to an OD600 of ≈0.6 (supplemented with the appropriate lysine; 40 μg/ml). Cultures were pelleted using an ultracentrifuge (8,000 g, 40 minutes) at 4°C and resuspended in chilled 20 ml lysis buffer containing 1% (w/v) n-dodecyl-beta-maltoside. The pellets could also be stored at -80°C for later use. Cells were then lysed with a Cell Disruptor (CF Range, Constant Systems Ltd., UK) and following removal of debris by centrifugation, concentrated to ~150 mg/ml using Amicon Ultra-15 centrifugation units (3kDa MWCO, Millipore). This provided about 600 μl of lysate, suitable for about six 80 μl DNA affinity purifications. Total protein concentration was assayed using the Bradford reagent (Sigma-Aldrich, St. Louis, MO). Following adjustment of protein concentration, sheared salmon sperm competitor DNA was added to the lysates (1 μg/ml; Life Technologies, Carlsbad, CA) and incubated for 10 minutes at 4°C. Finally, following centrifugation at 14,000 g to remove insoluble matter, the cell lysates were incubated for 1 hour with washed magnetic beads that contained no tethered DNA (0.5 mg beads per 100 μl lysate). Lysates were then either placed on ice or stored at 4°C prior to use.

Before performing affinity chromatography experiments, we also confirmed heavy lysine was being incorporated. Here, MG1655 $\Delta lysA$::kan cells from an overnight M9 minimal media culture were diluted 1:200 and 1:1,000, and grown in 1 ml M9 minimal media supplemented with 40 μg/ml heavy lysine. Following approximately 7 and 10 cell divisions, cells were resuspended in lysis buffer (50 mM HEPES pH 7.5, 70 mM potassium acetate, 5 mM magnesium acetate, 0.2% (w/v) n-dodecyl-beta-D-maltoside, Roche protease inhibitor cOmplete tablet) and lysed by performing 10 freeze-thaw cycles with dry ice. Cellular debris was removed by centrifugation at 14000 g at 4°C on a tabletop centrifuge. Finally cellular lysates were prepared for mass spectrometry by in-solution digestion with endoproteinase Lys-C (Promega, Madison, WI). Digestion was performed as described elsewhere [110] and labeling of the heavy isotope was confirmed by mass spectrometry measurement. In addition, we also characterized the SILAC enrichment ratio measurement by directly combining measurements from heavy and light lysates over a range from 0.1:1 to

1,000:1 heavy:light (see Supplemental Section 3.9).

Preparation of DNA-tethered magnetic beads

DNA affinity chromatography was performed by incubating cell lysate with magnetic beads (Dynabeads MyOne T1, ThermoFisher, Waltham, MA) containing tethered DNA. The DNA was tethered through a linkage between streptavidin on the beads and biotin on the DNA. Note that single-stranded DNA was purchased from Integrated DNA Technologies with the biotin modification on the 5' end of the oligonucleotide sense strand.

To begin preparation of tethered beads, DNA was suspended in annealing buffer (20 mM Tris-HCl, 10 mM MgCl2, 100 mM KCl) to 50 μM. Complementary strands were annealed by mixing 30 μL of the sense strand and 40 μL of the complement strand. Excess complement strand ensured all biotinylated-DNA would be in a double stranded form. Annealing was then performed using a thermocycler: 90°C for 5 minutes, gradient from 90°C to 65°C @ 0.1C /sec, incubated for 10 minutes at 65°C and allowed to return to room temperature on the thermocycler. Prior to attaching DNA, 150 μL beads were washed twice with 600 μL TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA) and then twice with DW buffer (20 mM Tris-HC pH 8.0, 2 M NaCl, 0.5 mM EDTA [21]). Approximately 640 pmol of DNA were then diluted to 600 μL in DW Buffer and incubated with the washed beads overnight at 4°C and on a rotatory wheel. Bound DNA was measured by determining the DNA concentration before and after incubation with beads using a NanoDrop (ThermoFisher, Waltham, MA). Finally, beads were washed once with 600 μL TE buffer and three washes of 600 μL DW buffer, and resuspended in 150 μL DW buffer.

DNA affinity chromatography

DNA affinity chromatography was performed by incubating cell lysate with magnetic beads (Dynabeads MyOne T1, ThermoFisher, Waltham, MA) containing tethered DNA. The DNA was tethered through a linkage between streptavidin on the beads and biotin on the DNA. Single-stranded DNA was purchased from Integrated DNA Technologies with the biotin modification on the 5' end of the oligonucleotide sense strand. Prior to DNA affinity purification the DNA tethered beads were incubated with blocking buffer (20 mM Hepes, pH 7.9, 0.05 mg/ml BSA, 0.05 mg/ml glycogen, 0.3 M KCl, 2.5 mM DTT, 5 mg/ml polyvinylpyrrolidone, 0.02% (w/v) n-dodeyl-

β-D-maltoside; about 1.3 ml/mg beads [21]) for one hour at 4°C for passivation. Excess blocking buffer was removed by washing the beads twice with 600 μL lysis buffer.

Cell lysates were incubated on a rotating wheel with the DNA tethered beads overnight at 4°C. Beads were recovered with a magnet and washed three times using an equivalent volume of lysis buffer. The beads were then washed once more, but with NEB Buffer 3.1 (New England Biolabs, MA, USA). Both purifications (with the target DNA and reference control) were combined by resuspending in 50 μL NEB Buffer 3.1, and 10 μl of the restriction enzyme PstI (100,000 units/ml, New England Biolabs) was added and incubated for 1.5 hours at 25°C. PstI cleaves the sequence CTGCAG, which was included between the biotin label and binding site sequence, allowing the DNA to be released from the magnetic beads. The beads were then removed and the samples prepared for mass spectrometry by in-gel digestion with endoproteinase Lys-C.

Note that in general, proteins were purified from a heavy lysate using DNA containing the target binding site sequence, while devoting the light lysate to a control DNA sequence. However, for our LacI and RelBE experiments, we also performed the alternative scenario, using the target sequence with the light lysate, and did not observe notable differences.

In-gel digestion of purified protein samples

Protein samples were diluted with 4x SDS-PAGE sample buffer and incubated for five minutes at 95°C and loaded on a SDS-PAGE gel (Any kD Mini-PROTEAN TGX Precast Protein Gels, 10-well , 50 µl; BioRad, CA, USA). Electrophoresis was performed for 45-55 minutes (200V) to provide 1-D size separation, and stained using the Colloidal Blue Staining Kit (ThermoFisher Scientific, MA, USA) for visualization. Destaining was performed with 100 mM ammonium bicarbonate, and the gel was cut into four sections, each of which was cut into roughly 1 mm pieces for in-gel digestion. The gel pieces were reduced, alkylated, and digested by endoproteinase Lys-C overnight at 37°C. This enzymatically cleaves proteins after lysine residues and is necessary for determining whether detected peptides are from the light or heavy lysine labeled purification. Digested peptides were then extracted from the gel and lyophilized. The peptide samples were further purified using StageTips to remove residual salts [111] and re-suspended in 0.2% formic acid.

LC-MS/MS method details

Liquid chromatography tandem-mass spectrometry (LC-MS/MS) experiments were carried out as previously described [79].

The LacI target purification experiments were performed on a nanoflow LC system, EASY-nLC II coupled to a hybrid linear ion trap Orbitrap Classic mass spectrometer equipped with a Nanospray Flex Ion Source (Thermo Fisher Scientific). The in-gel digested peptides were directly loaded at a flow rate of 500 nl/min onto a 16-cm analytical HPLC column (75 μm ID) packed in-house with ReproSil-Pur C18AQ 3 μm resin (120 Å pore size, Dr. Maisch, Ammerbuch, Germany). The column was enclosed in a column heater operating at 45°C. After 30 min of loading time, the peptides were separated in a solvent gradient at a flow rate of 350 nl/min. The gradient was as follows: 0–30% B (80 min), and 100% B (10 min). The solvent A consisted of 97.8% H2O, 2% ACN, and 0.2% formic acid and solvent B consisted of 19.8% H2O, 80% ACN, and 0.2% formic acid. The Orbitrap was operated in data-dependent acquisition mode to automatically alternate between a full scan (m/z=400–1600) in the Orbitrap (resolution 100,000) and subsequent 15 CID MS/MS scans (Top 15 method) in the linear ion trap. Collision induced dissociation (CID) was performed at normalized collision energy of 35% and 30 msec of activation time.

All other measurements were performed on a hybrid ion trap-Orbitrap Elite mass spectrometer (Thermo Fisher Scientific), which provided greater detection sensitivity and other fragmentation techniques as described. The Orbitrap was operated in data-dependent acquisition mode to automatically alternate between a full scan (m/z=400–1,800) in the Orbitrap (resolution 120,000) and subsequent 5 MS/MS scans also acquired in Orbitrap with 15,000 resolution. The MS/MS spectra were acquired for the top 5 ions alternating between higher collision dissociation (HCD) and electron transfer dissociation (ETD) fragmentations that are well suited for higher charge peptides. Higher collision dissociation was performed at a normalized collision energy of 30% and electron transfer dissociation reaction time was set to 100 msec. The analytical column for this instrument was a PicoFrit column (New Objective, Woburn, MA) packed in house with ReproSil-Pur C18AQ 1.9 μm resin (120Å pore size, Dr. Maisch, Ammerbuch, Germany) and the column was heated to 60°C. The peptides were separated either with a 90 or 60 min gradient (0-30% B in 90 min or 0-30% B in 60 min) at a flow rate of 220 nL/min.

Mass spectrometry data processing

Thermo RAW files were processed using MaxQuant (v. 1.5.3.30) [80, 112]. Spectra were searched against the UniProt *E. coli* K-12 database (4318 sequences) as well as a contaminant database (256 sequences). Additional details are provided in the supplemental methods. Precursor ion mass tolerance was 4.5 ppm after recalibration by MaxQuant. Fragment ion mass tolerance was 20 ppm for high-resolution HCD and ETD spectra, and 0.5 Da for low-resolution CID spectra. Variable modifications included oxidation of methionine and protein N-terminal acetylation. Carboxyamidomethylation of cysteine was specified as a fixed modification. LysC was specified as the digestion enzyme and up to two missed cleavages were allowed. A decoy database was generated by MaxQuant and used to set a score threshold so that the false discovery rate was less than 1% at both the peptide and protein level. For all experiments match between runs and re- quantify were enabled. One evidence ratio per replicate per protein was required for quantitation.

To calculate the overall protein ratio, the non-normalized protein replicate ratios were log transformed and then shifted so that the median protein log ratio within each replicate was zero (i.e., the median protein ratio was 1:1). The overall experimental log ratio was then calculated from the average of the replicate ratios. Proteins were considered if they were known to be transcription factors, or predicted to bind DNA (using gene ontology term GO:0003677, for DNA-binding in BioCyc).

BIBLIOGRAPHY

[1] Socorro Gama-Castro, Heladia Salgado, Alberto Santos-Zavaleta, Daniela Ledezma-Tejeida, Luis Muñiz-Rascado, Jair Santiago García-Sotelo, Kevin Alquicira-Hernández, Irma Martínez-Flores, Lucia Pannier, Jaime Abraham Castro-Mondragón, Alejandra Medina-Rivera, Hilda Solano-Lira, César Bonavides-Martínez, Ernesto Pérez-Rueda, Shirley Alquicira-Hernández, Liliana Porrón-Sotelo, Alejandra López-Fuentes, Anastasia Hernández-Koutoucheva, Víctor Del Moral-Chávez, Fabio Rinaldi, and Julio Collado-Vides. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 44(D1):D133–D143, 2016.

[2] Ingrid M. Keseler, Amanda Mackie, Martin Peralta-Gil, Alberto Santos-Zavaleta, Socorro Gama-Castro, César Bonavides-Martinez, Carol Fulcher, Araceli M. Huerta, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Luis Muñiz-Rascado, Quang Ong, Suzanne Paley, Imke Schroeder, Alexander G. Shearer, Pallavi Subhraveti, Mike Travers, Deepika Weerasinghe, Verena Weiss, Julio Collado-Vides, Robert P. Gunsalus, Ian Paulsen, and Peter D. Karp. EcoCyc: Fusing model organism databases with systems biology. *Nucleic Acids Research*, 41(D1):D605–D612, 2013.

[3] Richard Münch, Karsten Hiller, Heiko Barg, Dana Heldt, Simone Linz, Edgar Wingender, and Dieter Jahn. PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Research*, 31(1):266–269, 2003.

[4] Michael J. Cipriano, Pavel N. Novichkov, Alexey E. Kazakov, Dmitry A. Rodionov, Adam P. Arkin, Mikhail S. Gelfand, and Inna Dubchak. RegTransBase – a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. *BMC Genomics*, 14(1):213–221, 2013.

[5] Sefa Kılıç, Elliot R. White, Dinara M. Sagitova, Joseph P. Cornish, and Ivan Erill. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Research*, 42(D1):D156–D160, 2013.

[6] David C. Grainger, Douglas Hurd, Marcus Harrison, Jolyon Holdstock, and Stephen J. W. Busby. Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proceedings of the National Academy of Sciences*, 102(49):17693–17698, 2005.

[7] Tiffany Vora, Alison K. Hottes, and Saeed Tavazoie. Protein occupancy landscape of a bacterial genome. *Molecular Cell*, 35(2):247–253, 2009.

[8] Richard P. Bonocora and Joseph T. Wade. *ChIP-Seq for genome-scale analysis of bacterial DNA-binding proteins*. New York, Humana Press, 2015.

[9] Dongling Zheng, Chrystala Constantinidou, Jon L. Hobman, and Stephen D. Minchin. Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. *Nucleic Acids Research*, 32(19):5874–5893, 2004.

[10] Shivani S. Singh, Navjot Singh, Richard P. Bonocora, Devon M. Fitzgerald, Joseph T. Wade, and David C. Grainger. Widespread suppression of intragenic transcription initiation by H-NS. *Genes and Development*, 28(3):214–219, 2014.

[11] Joseph T. Wade. ChIP-Seq for genomic-scale analysis of bacterial DNA-binding proteins. *Prokaryotic Systems Biology*, 883(Chapter 7):119–134, 2015.

[12] Stephen D. Minchin and Stephen J. W. Busby. Analysis of mechanisms of activation and repression at bacterial promoters. *Methods*, 47(1):6–12, 2009.

[13] Justin B. Kinney, Anand Murugan, Curtis G. Callan, and Edward C. Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*, 107(20):9158–9163, 2010.

[14] Alexandre Melnikov, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, Andreas Gnirke, Curtis G. Callan, Justin B. Kinney, Manolis Kellis, Eric S. Lander, and Tarjei S. Mikkelsen. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3):271–277, 2012.

[15] P. Kheradpour, J. Ernst, A. Melnikov, P. Rogov, L. Wang, X. Zhang, J. Alston, T. S. Mikkelsen, and M. Kellis. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, 23(5):800–811, 2013.

[16] Rupali P. Patwardhan, Joseph B. Hiatt, Daniela M. Witten, Mee J. Kim, Robin P. Smith, Dalit May, Choli Lee, Jennifer M. Andrie, Su-In Lee, Gregory M. Cooper, Nadav Ahituv, Len A. Pennacchio, and Jay Shendure. Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nature Biotechnology*, 30(3):265–270, 2012.

[17] Eilon Sharon, Yael Kalma, Ayala Sharp, Tali Raveh-Sadka, Michal Levo, Danny Zeevi, Leeat Keren, Zohar Yakhini, Adina Weinberger, and Eran Segal. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, 30(6):521–530, 2012.

[18] Sriram Kosuri, Daniel B. Goodman, Guillaume Cambray, Vivek K. Mutalik, Yuan Gao, Adam P. Arkin, Drew Endy, and George M. Church. Composability of regulatory sequences controlling transcription and translation

in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 110(34):14024–14029, 2013.

[19] Brett B. Maricque, Joseph D. Dougherty, and Barak A. Cohen. A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Research*, 45(4):e16–e16, 2017.

[20] Charles P. Fulco, Mathias Munschauer, Rockwell Anyoha, Glen Munson, Sharon R. Grossman, Elizabeth M Perez, Michael Kane, Brian Cleary, Eric S. Lander, and Jesse M. Engreitz. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science*, 354(6313):769–773, 2016.

[21] Gerhard Mittler, Falk Butter, and Matthias Mann. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Research*, 19(2):284–293, 2009.

[22] Hamid Mirzaei, Theo A. Knijnenburg, Bong Kim, Max Robinson, Paola Picotti, Gregory W. Carter, Song Li, David J. Dilworth, Jimmy K. Eng, John D. Aitchison, Ilya Shmulevich, Timothy Galitski, Ruedi Aebersold, and Jeffrey Ranish. Systematic measurement of transcription factor-DNA interactions by targeted mass spectrometry identifies candidate gene regulatory proteins. *Proceedings of the National Academy of Sciences*, 110(9):3645–3650, 2013.

[23] R. Lutz and H. Bujard. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research*, 25(6):1203–1210, 1997.

[24] Ville Mustonen, Justin Kinney, Curtis G. Callan, and Michael Lassig. Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. *Proceedings of the National Academy of Sciences*, 105(34):12376–12381, 2008.

[25] William T. Ireland and Justin B. Kinney. MPAthic: quantitative modeling of sequence-function relationships for massively parallel assays. *bioRxiv*, page 054676, 2016.

[26] Thomas D. Schneider and R. Michael Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, 1990.

[27] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular Cell Proteomics*, 1(5):376–386, 2002.

[28] Michael J. Kerner, Dean J. Naylor, Yasushi Ishihama, Tobias Maier, Hung-Chun Chang, Anna P. Stines, Costa Georgopoulos, Dmitrij Frishman, Manajit Hayer-Hartl, Matthias Mann, and F. Ulrich Hartl. Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli*. *Cell*, 122(2):209–220, 2005.

[29] Giulia Calloni, Taotao Chen, Sonya M. Schermann, Hung-Chun Chang, Pierre Genevaux, Federico Agostini, Gian Gaetano Tartaglia, Manajit Hayer-Hartl, and F. Ulrich Hartl. DnaK functions as a central hub in the *E. coli* chaperone network. *Cell Reports*, 1(3):251–264, 2012.

[30] Boumediene Soufi and Boris Macek. Stable Isotope Labeling by Amino Acids Applied to Bacterial Cell Culture. In *Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)*, pages 9–22. Humana Press, New York, NY, New York, NY, 2014.

[31] Stefan Oehler, E. R. Eismann, Helmut Krämer, and Benno Müller-Hill. The three operators of the *lac* operon cooperate in repression. *The EMBO Journal*, 9(4):973–979, 1990.

[32] Kenn Gerdes, Susanne K. Christensen, and Anders Løbner-Oleson. Prokaryotic toxin–antitoxin stress response loci. *Nature Reviews Microbiology*, 2(5):371–382, 2005.

[33] Michael N. Alekshun and Stuart B. Levy. Regulation of chromosomally mediated multiple antibiotic resistance: the *mar* regulon. *Journal of Molecular Biology*, 41(10):2067–2075, 1997.

[34] Hernan G. Garcia and Rob Phillips. Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences*, 108(29):12173–12178, 2011.

[35] Etienne Maisonneuve and Kenn Gerdes. Molecular mechanisms underlying bacterial persisters. *Cell*, 157(3):539–548, 2014.

[36] Martin Overgaard, Jonas Borch, and Kenn Gerdes. Bacterial toxin RelE: A highly efficient ribonuclease with exquisite substrate specificity using atypical catalytic residues. *Biochemistry*, 52(48):8633–8642, 2013.

[37] Martin Overgaard, Jonas Borch, and Kenn Gerdes. RelB and RelE of *Escherichia coli* form a tight complex that represses transcription via the ribbon–helix–helix motif in RelB. *Journal of Molecular Biology*, 394(2):183–196, 2009.

[38] Robert G. Martin and Judah L. Rosner. Fis, an accessorial factor for transcriptional activation of the mar (multiple antibiotic resistance) promoter of *Escherichia coli* in the presence of the activator MarA, SoxS, or Rob. *Journal of Bacteriology*, 179(23):7410–7419, 1997.

[39] Asuncion S. Seoane and Stuart B. Levy. Characterization of MarR, the repressor of the multiple antibiotic resistance (mar) operon in Escherichia coli. *Journal of Bacteriology*, 177(12):3414–3419, 1995.

[40] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S. Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–635, 2014.

[41] Thomas Kuhlman, Zhongge Zhang, Milton H. Saier, and Terence Hwa. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 104(14):6043–6048, 2007.

[42] Guang-Yao Li, Yonglong Zhang, Masayori Inouye, and Mitsuhiko Ikura. Structural mechanism of transcriptional autorepression of the *Escherichia coli* RelB/RelE antitoxin/toxin module. *Journal of Molecular Biology*, 380(1):107–119, 2008.

[43] Martin Overgaard, Jonas Borch, Mikkel G. Jørgensen, and Kenn Gerdes. Messenger RNA interferase RelE controls relBE transcription by conditional cooperativity. *Molecular Microbiology*, 69(4):841–857, 2008.

[44] Daniel Marbach, James C. Costello, Robert Kueffner, Robert J. Vega, Nicole M.and Prill, Diogo M. Camacho, Kyle R. Allison, Manolis Kellis, James J. Collins, Gustavo Stolovitzky, and DREAM5 Consortium. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, 2012.

[45] Alexander Schmidt, Karl Kochanowski, Silke Vedelaar, Erik Ahrne, Benjamin Volkmer, Luciano Callipo, Kevin Knoops, Manuel Bauer, Ruedi Aebersold, and Matthias Heinemann. The quantitative and condition-dependent *Escherichia coli* proteome. *Nature Biotechnology*, 34:104–111, 2016.

[46] Ronda J. Rolfes. Regulation of purine nucleotide biosynthesis: in yeast and beyond. *Biochemical Society Transactions*, 34(5):786–790, 2006.

[47] Byung-Kwan Cho, Stephen A. Federowicz, Mallory Embree, Young-Seoub Park, Donghyuk Kim, and Bernhard Palsson. The PurR regulon in *Escherichia coli* K-12 MG1655. *Nucleic Acids Research*, 39(15):6456–6464, 2011.

[48] Mariza R. Lomba, Ana T. Vasconcelos, Ana Beatriz F. Pacheco, and Darcy F. Almeida. Identification of *yebG* as a DNA damage-inducible *Escherichia coli* gene. *FEMS Microbiology Ecology*, 156(1):119–122, 1997.

[49] Joseph T. Wade, Nikos B. Reppas, George M. Church, and Kevin Struhl. Genomic analysis of LexA binding reveals the permissive nature of the *Escherichia coli* genome and identifies unconventional target sites. *Genes and Development*, 19(21):2619–2630, 2005.

[50] Kang Y. Choi and Howard Zalkin. Structural characterization and corepressor binding of the *Escherichia coli* purine repressor. *Journal of Bacteriology*, 174(19):6207–6214, 1992.

[51] Lacramioara Bintu, Nicolas E. Buchler, Hernan G. Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips. Transcriptional regulation by the numbers: Models. *Current Opinion in Genetics and Development*, 15(2):116–124, 2005.

[52] Gurinder S. Atwal and Justin B. Kinney. Learning quantitative sequence-function relationships from massively parallel experiments. *Journal of Statistical Physics*, 162(5):1203–1243, 2016.

[53] Sukgil Song and Chankyu Park. Organization and regulation of the D-xylose operons in *Escherichia coli* K-12: XylR acts as a transcriptional activator. *Journal of Bacteriology*, 179(22):7025–7032, 1997.

[54] Douglas F. Browning and Stephen J. W. Busby. Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, 14(10):638–650, 2016.

[55] Olga N. Laikova, Andrey A. Mironov, and Mikhail S. Gelfand. Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria. *FEMS Microbiology Ecology*, 205(2):315–322, 2001.

[56] Ronald A. Cooper. The utilisation of D-galactonate and D-2-oxo-3-deoxygalactonate by *Escherichia coli* K-12. Biochemical and genetical studies. *Arcives of Microbiology*, 1(118):199–206, 1978.

[57] Brandon Ho, Anastasia Baryshnikova, and Grant W. Brown. Unification of protein abundance datasets yields a quantitative *Saccharomyces cerevisiae* proteome. *Cell Systems*, 6:1–14, 2018.

[58] Thomas E. Kuhlman and Edward C. Cox. Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Research*, 38(6):e92–e92, 2010.

[59] Huibin Zhang, Teodorus T. Susanto, Yue Wan, and Swaine L. Chen. Comprehensive mutagenesis of the fimS promoter regulatory switch reveals novel regulation of type 1 pili in uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences*, 113(15):4182–4187, 2016.

[60] Guillaume Urtecho, Arielle D. Tripp, Kimberly Insigne, Hwangbeom Kim, and Sriram Kosuri. Systematic dissection of sequence elements controlling σ^{70} promoters using a genomically-encoded multiplexed reporter assay in *E. coli*. *Biochemistry*, 2018.

[61] Irina O. Vvedenskaya, Yuanchao Zhang, Seth R. Goldman, Anna Valenti, Valeria Visone, Deanne M. Taylor, Richard H. Ebright, and Bryce E. Nickels. Massively systematic transcript end readout, "MASTER": Transcription start site selection, transcriptional slippage, and transcript yields. *Molecular Cell*, 60(6):953–965, 2015.

[62] Andrew Thompson, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, 75(8):1895–1904, 2003.

[63] Philip L. Ross, Yulin N. Huang, Jason N. Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, Sasi Pillai, Subhakar Dey, Scott Daniels, Subhasish Purkayastha, Peter Juhasz, Stephen Martin, Michael Bartlet-Jones, Feng He, Allan Jacobson, and Darryl J. Pappin. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular Cell Proteomics*, 3(12):1154–1169, 2004.

[64] Brian K. Erickson, Christopher M. Rose, Craig R. Braun, Alison R. Erickson, Jeffrey Knott, Graeme C. McAlister, Martin Wuhr, Joao A. Paulo, Robert A. Everley, and Steven P. Gygi. A strategy to combine sample multiplexing with targeted proteomics assays for high-throughput protein signature characterization. *Molecular Cell*, 65(2):361–370, 2017.

[65] Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347–355, 2016.

[66] Nina C. Hubner, Luan N. Nguyen, Nadine C. Hornig, and Hendrik G. Stunnenberg. A quantitative proteomics tool to identify DNA-protein interactions in primary cells or blood. *Journal of Proteome Research*, 14(2):1315–1329, 2014.

[67] Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A. Datsenko, Masaru Tomita, Barry L. Wanner, and Hirotada Mori. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Molecular Systems Biology*, 2(1):2006.0008, 2006.

[68] Byoung-Mo Koo, George Kritikos, Jeremiah D. Farelli, Horia Todor, Kenneth Tong, Harvey Kimsey, Ilan Wapinski, Marco Galardini, Angelo Cabal, Jason M. Peters, Anna-Barbara Hachmann, David Z. Rudner, Karen N. Allen, Athanasios Typas, and Carol A. Gross. Construction and analysis of two genome-scale deletion libraries for *Bacillus subtilis*. *Cell Systems*, 4(3):291–305.e7, 2017.

[69] Véronique de Berardinis, David Vallenet, Vanina Castelli, Marielle Besnard, Agnès Pinet, Corinne Cruaud, Sumitta Samair, Christophe Lechaplais, Gabor Gyapay, Céline Richez, Maxime Durot, Annett Kreimeyer, François Le Fèvre, Vincent Schächter, Valérie Pezo, Volker Döring, Claude Scarpelli, Claudine Médigue, Georges N Cohen, Philippe Marlière, Marcel Salanoubat, and Jean Weissenbach. A complete collection of single-gene deletion mutants of *Acinetobacter baylyi* ADP1. *Molecular Systems Biology*, 4(1):174–154, 2008.

[70] Steffen Porwollik, Carlos A. Santiviago, Pui Cheng, Fred Long, Prerak Desai, Jennifer Fredlund, Shabarinath Srikumar, Cecilia A. Silva, Weiping Chu, Xin Chen, Rocío Canals, M. Megan Reynolds, Lydia Bogomolnaya, Christine Shields, Ping Cui, Jinbai Guo, Yi Zheng, Tiana Endicott-Yazdani, Hee-Jeong Yang, Aimee Maple, Yury Ragoza, Carlos J. Blondel, Camila Valenzuela, Helene Andrews-Polymenis, and Michael McClelland. Defined single-gene and multi-gene deletion mutant collections in *Salmonella enterica* sv Typhimurium. *PLoS ONE*, 9(7):e99820, 2014.

[71] Ping Xu, Xiuchun Ge, Lei Chen, Xiaojing Wang, Yuetan Dou, Jerry Z. Xu, Jenishkumar R. Patel, Victoria Stone, My Trinh, Karra Evans, Todd Kitten, Danail Bonchev, and Gregory A. Buck. Genome-wide essential gene identification in *Streptococcus sanguinis*. *Scientific Reports*, 1(1):4555, 2011.

[72] Nicole T. Liberati, Jonathan M. Urbach, Sachiko Miyata, Daniel G. Lee, Eliana Drenkard, Gang Wu, Jacinto Villanueva, Tao Wei, and Frederick M Ausubel. An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proceedings of the National Academy of Sciences*, 103(8):2833–2838, 2006.

[73] Matthew H. Larson, Luke A. Gilbert, Xiaowo Wang, Wendell A. Lim, Jonathan S. Weissman, and Lei S. Qi. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature Protocols*, 8(11):2180–2196, 2013.

[74] Gina C. Gordon, Travis C. Korosh, Jeffrey C. Cameron, Andrew L. Markley, Matthew B. Begemann, and Brian F. Pfleger. CRISPR interference as a titratable, trans-acting regulatory tool for metabolic engineering in the cyanobacterium *Synechococcus* sp. strain PCC 7002. *Metabolic Engineering*, 38:170–179, 2016.

[75] Michael Lässig. From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics*, 8(Suppl 6):S7–21, 2007.

[76] Alexander J. Stewart and Joshua B. Plotkin. Why transcription factor binding sites are ten nucleotides long. *Genetics*, 192(3):973–985, 2012.

[77] Andrey Feklístov, Brian D. Sharon, Seth A. Darst, and Carol A. Gross. Bacterial sigma factors: A historical, structural, and genomic perspective. *Annual Review of Microbiology*, 68(1):357–376, 2014.

[78] Kirill A. Datsenko and Barry L. Wanner. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences*, 97(12):6640–6645, 2000.

[79] Anastasia Kalli and Sonja Hess. Effect of mass spectrometric parameters on peptide and protein identification rates for shotgun proteomic experiments on an LTQ-orbitrap mass analyzer. *Proteomics*, 12(1):21–31, 2011.

[80] Jürgen Cox, Ivan Matic, Maximiliane Hilger, Nagarjuna Nagaraj, Matthias Selbach, Jesper V. Olsen, and Matthias Mann. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nature Protocols*, 4(5):698–705, 2009.

[81] Shujiro Okuda, Yu Watanabe, Yuki Moriya, Shin Kawano, Tadashi Yamamoto, Masaki Matsumoto, Tomoyo Takami, Daiki Kobayashi, Norie Araki, Akiyasu C. Yoshizawa, Tsuyoshi Tabata, Naoyuki Sugiyama, Susumu Goto, and Yasushi Ishihama. jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Research*, 45(D1):D1107–D1111, 2017.

[82] M. H. Irani, L. Orosz, and S. Adhya. A control element within a structural gene: The gal operon of Escherichia coli. *Cell*, 32(3):783–788, 1983.

[83] Szabolcs Semsey, Sandeep Krishna, Kim Sneppen, and Sankar Adhya. Signal integration in the galactose network of Escherichia coli. *Molecular Microbiology*, 65(2):465–476, July 2007.

[84] Keishin Nishida, Martin C. Frith, and Kenta Nakai. Pseudocounts for transcription factor binding sites. *Nucleic Acids Research*, 37(3):939–944, 2009.

[85] Xuhua Xia. Position weight matrix, Gibbs sampler, and the associated significance tests in motif characterization and prediction. *Scientifica*, 2012(1):1–15, November 2012.

[86] Gary D. Stormo. DNA binding sites: Representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[87] Otto G. Berg and Peter H. von Hippel. Selection of DNA binding sites by regulatory proteins. *Journal of Molecular Biology*, 193(4):723–743, 1987.

[88] Thomas D. Schneider, Gary D. Stormo, Larry Gold, and Andrzej Ehrenfeucht. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, 188(3):415–431, 1986.

[89] Shobhit Gupta, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, 2007.

[90] Franz M. Weinert, Robert C. Brewster, Mattias Rydenfelt, Rob Phillips, and Willem K Kegel. Scaling of gene expression with transcription-factor fugacity. *Physical Review Letters*, 113(25):258101, 2014.

[91] Uri Moran, Rob Phillips, and Ron Milo. SnapShot: Key numbers in biology. *Cell*, 141(7):1262–1262.e1, 2010.

[92] Shao-En Ong and Matthias Mann. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nature Protocols*, 1(6):2650–2660, January 2007.

[93] Mattias Rydenfelt, Hernan G. Garcia, Robert Sidney Cox, III, and Rob Phillips. The influence of promoter architectures and regulatory motifs on gene expression in Escherichia coli. *PLoS ONE*, 9(12):e114347, December 2014.

[94] Mads T. Bonde, Sriram Kosuri, Hans J. Genee, Kira Sarup-Lytzen, George M. Church, Morten O. A. Sommer, and Harris H. Wang. Direct mutagenesis of thousands of genomic targets using microarray-derived oligonucleotides. *ACS Synthetic Biology*, 4(1):17–22, January 2015.

[95] Julia Rohlhill, Nicholas R. Sandoval, and Eleftherios T Papoutsakis. Sort-Seq approach to engineering a formaldehyde-inducible promoter for dynamically regulated *Escherichia coli* growth on methanol. *ACS Synthetic Biology*, 6:1584–1595, 2017.

[96] Howard M. Salis, Ethan A. Mirsky, and Christopher A. Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, 27(10):946–950, oct 2009.

[97] Joseph N. Zadeh, Conrad D. Steenberg, Justin S. Bois, Brian R. Wolfe, Marshall B. Pierce, Asif R. Khan, Robert M. Dirks, and Niles A. Pierce. NUPACK: Analysis and design of nucleic acid systems. *Journal of Computational Chemistry*, 32(1):170–173, 2011.

[98] Daniel L. Jones, Robert C. Brewster, and Rob Phillips. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216):1533–1536, 2014.

[99] Alessandro Treves and Stefano Panzeri. The upward bias in measures of information derived from limited data samples. *Neural Computation*, 7(2):399–407, March 1995.

[100] Justin B. Kinney, Gašper Tkačik, and Curtis G. Callan. Precise physical models of protein-DNA interaction from high-throughput data. *Proceedings of the National Academy of Sciences*, 104(2):501–506, 2007.

[101] Anand Patil, David Huard, and Christopher J. Fonnesbeck. PyMC: Bayesian stochastic modelling in python. *Journal of Statistical Software*, 35(4):1–811, July 2010.

[102] Devinderjit Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. (Oxford, Oxford University Press), 2006.

[103] Justin B. Kinney and Gurinder S. Atwal. Parametric inference in the large data limit using maximally informative models. *Neural Computation*, 26(4):637–653, 2014.

[104] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306–312, 2013.

[105] C. Ushida and H. Aiba. Helical phase dependent action of CRP: effect of the distance between the CRP site and the -35 region on promoter activity. *Nucleic Acids Research*, 18(21):6325–6330, November 1990.

[106] K. Gaston, A. Bell, A. Kolb, H. Buc, and S. Busby. Stringent spacing requirements for transcription activation by CRP. *Cell*, 62(4):733–743, August 1990.

[107] M. Razo-Mejia, J. Q. Boedicker, D. Jones, A. DeLuna, J. B. Kinney, and R. Phillips. Comparison of the theoretical and real-world evolutionary potential of a genetic circuit. *Phys Biol*, 11(2):026005, April 2014.

[108] Daniel G. Gibson, Lei Young, Ray-Yuan Chuang, J. Craig Venter, Clyde A. Hutchison, and Hamilton O. Smith. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, 6(5):343–345, April 2009.

[109] T. Magoc and S. L. Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, October 2011.

[110] Jacek R. Wisniewski, Alexandre Zougman, Nagarjuna Nagaraj, and Matthias Mann. Universal sample preparation method for proteome analysis. *Nature Methods*, 6(5):359–362, April 2009.

[111] J. Rappsilber, M. Mann, and Y. Ishihama. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature Protocols*, 2(8):1896–1906, 2007.

[112] Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, December 2008.

Chapter 4

MAPPING DNA SEQUENCE TO TRANSCRIPTION FACTOR BINDING ENERGY *IN VIVO*

This work was performed in collaboration with N. M. Belliveau, W. T. Ireland, J. B. Kinney, and R. Phillips.

Author contribution note: for this chapter, I (SB) performed Sort-Seq sample processing, strain construction, fold-change measurements, data analysis, and was the primary writer for the manuscript. This work is in preparation for publication in a peer-reviewed journal.

4.1 Introduction

High-throughput sequencing allows us to sequence the genome of nearly any species at will. The amount of genomic data available is already enormous and will only continue to grow. However, this mass of data is largely uninformative without appropriate methods of analyzing it. Despite decades of research, much genomic data still defies our efforts to interpret it. It is particularly challenging to interpret non-coding DNA such as intergenic regulatory regions. We can infer the locations of some transcription start sites and transcription factor binding sites, but these inferences tell us little about the functional role of these putative sites. In order to better interpret these types of sequences, we need a better understanding of how sequence elements control gene expression. An important avenue for developing this level of understanding is to propose models that map sequence to function and perform experiments that test these models.

Over half of the genes in *E. coli*, which is arguably the best-understood model organism, lack any regulatory annotation (see RegulonDB [1]). Those operons whose regulation is well described (e.g. the *lac*, *rel*, and *mar* operons [2–4]) required decades of work, often involving laborious genetic and biochemical experiments [5]. A wide variety of new techniques have been proposed and implemented to simplify the process of determining how a gene is regulated. Chromatin immunoprecipitation (ChIP) based methods such as ChIP-chip and ChIP-seq make it possible to determine the genome-wide binding locations of individual transcription factors of interest. Massively parallel reporter assays (MPRAs) have made it possible to read

out transcription factor binding position and occupancy *in vivo* with base-pair resolution, and provide a means for analyzing non-binding features such as "insulator" sequences [6–8]. *In vitro* methods based on protein-binding microarrays [9], SELEX [10–12], MITOMI [13–15], and binding assays performed in high-throughput sequencing flow cells [16, 17] have made it possible to measure transcription factor affinity to a broad array of possible binding sites and can also account for features such as flanking sequences [15, 18, 19]. However, *in vitro* methods cannot fully account for the *in vivo* consequences of binding site context and interactions with other proteins. Current *in vivo* methods for measuring transcription factor binding affinities, such as bacterial one-hybrid [20, 21], require a restructuring of the promoter so that it no longer resembles its genomic counterpart. Additionally, efforts to computationally ascertain the locations of transcription factor binding sites frequently produce false positives [22, 23]. Furthermore, a common assumption underlying many of these methods is that transcription factor occupancy in the vicinity of a promoter implies regulation, but it has been shown that occupancy cannot always accurately predict the effect of a transcription factor on gene regulation [24, 25]. As these examples show, it remains challenging to integrate multiple aspects of transcription factor binding into a cohesive understanding of gene regulation.

Here we work to develop such a cohesive understanding by integrating rigorous thermodynamic modeling with *in vivo* transcription factor binding experiments. In Ref. [26], we showed that the MPRA Sort-Seq [27], combined with a simple linear model for protein-DNA binding specificity, can be used to accurately predict the binding energies of multiple RNAP binding site mutants, serving as a jumping off point for the use of such models as a quantitative tool in synthetic biology. Here we apply this technique to transcription factor binding sites in an effort to better understand how transcription factors interact with regulatory DNA under different conditions. Specifically, we use Sort-Seq to map sequence to binding energy for a repressor-operator interaction, and we rigorously characterize the variables that must be considered in order to obtain an accurate sequence-binding energy map. We then use our sequence-energy mapping to design a series of operators with a hierarchy of controlled binding energies measured in k_BT units. To demonstrate our control over these operators and their associated regulatory logic, we use these characterized binding sites to design a wide range of induction responses with different phenotypic properties such as leakiness, dynamic range and $[EC_{50}]$. Next, we focus our attention on the synergy between mutations in the amino acid sequence of transcription factors and their corresponding binding sites. Finally, we show the broader reach of these

results by exploring how binding site position and regulatory context can change the DNA-protein sequence specificity for multiple different transcription factors.

4.2 Results

Obtaining energy matrices using Sort-Seq

A major goal of this study was to show that one can use Sort-Seq to precisely map DNA sequence to binding energy for a transcription factor binding site, thus making it possible to predict and manipulate transcriptional activity *in vivo*. While numerous *in vitro* studies have successfully mapped sequence to affinity [9–11, 13, 14, 16, 17], and some *in vivo* studies have used methods such as bacterial one-hybrid to provide such mappings as well [20, 21], these studies are limited because they cannot be adapted to reflect the actual wild-type arrangement of regulatory elements, thus potentially missing vital regulatory information. Moreover, while position-weight matrices (PWMs) derived from genomic data have traditionally been used to ascertain *in vivo* sequence specificities, it can be difficult to convert these specificities into quantitative binding energy mappings due to the relatively small number of sequences that are used to generate these PWMs.

Sort-Seq has previously been shown to be a promising technique for mapping protein binding sequences to binding energies. In Ref. [26], binding energy predictions for RNAP were made from an energy matrix generated in Ref. [27] that used the wild-type *lac* promoter as a reference sequence (i.e. the sequence that was mutated to perform Sort-Seq). Here, we design experiments that use the Sort-Seq technique described in [27] with the specific intent of creating energy matrices with maximum predictive power (see Figure 4.1), and we test the predictions from these matrices against measured binding energies. We show that such predictive matrices can be produced for multiple transcription factors (e.g. XylR, PurR, and LacI) implicated in an array of regulatory architectures. To thoroughly test the accuracy of our predictive matrices, we begin with promoters that employ "simple repression," in which a repressor binds to an operator such that it occludes RNAP binding, thereby preventing transcription and repressing the gene [28]. As a model for how sequence-energy mappings might be used for transcription factor binding sites in simple repression architectures, we interrogate the binding specificity of the *lac* repressor (LacI). LacI was chosen for this role because it is well-characterized and has known binding sites in only one operon within the genome, making it an ideal choice for this kind of systematic and rigorous analysis. We create three distinct energy matrices in which each of the natural *lac* operators (O1, O2, or O3 [2]) acts as the reference sequence. Supplemental Section 4.5 lists the wild-type sequences for these simple repression constructs.

As described in Figure 4.1, to perform Sort-Seq we start by mutating the promoter at a rate of $\sim$ 10%. Here we mutate both the RNAP binding site and the operator, starting with either O1, O2, or O3 for the operator sequence. While our analysis focuses on the operators themselves, mutating the RNAP as well aids in model-fitting as described in Supplemental Section 4.6. We place the promoters upstream of a fluorescent reporter gene and create a plasmid library of these constructs. We transform this plasmid library into a population of *E. coli* in which *lacI* and *lacZYA* have been deleted, but *lacI* has been reintroduced to the genome with a synthetic RBS that allows us to precisely control the LacI copy number within the cell, as described in Ref. [29]. We require at least 10^6 transformants for each plasmid library to ensure sufficient library diversity. Then, we use fluorescence-activated cell sorting (FACS) to sort *E. coli* containing these plasmids into four bins based on their expression levels. We perform high-throughput sequencing on the libraries from each bin. We infer energy matrices that maximize the mutual information between sequence and expression bin (see Supplemental Sections 3.12 and 4.6 for details). We perform Bayesian parameter estimation using a Markov Chain Monte Carlo algorithm to determine the scaling factor that should be applied to the energy matrix to convert each position into k_BT energy units. We infer the scaling factor using the same data set that was used to infer the energy matrix, as the ideal scaling factor should maximize the mutual information between promoter sequence and gene expression (see Supplemental Section 4.7 for a comparison to other methods for obtaining the scaling factor). At this point, one can compute the expected binding energy of any operator mutant within several mutations of the reference sequence by simply adding together the energy values associated with each base in the operator mutant.

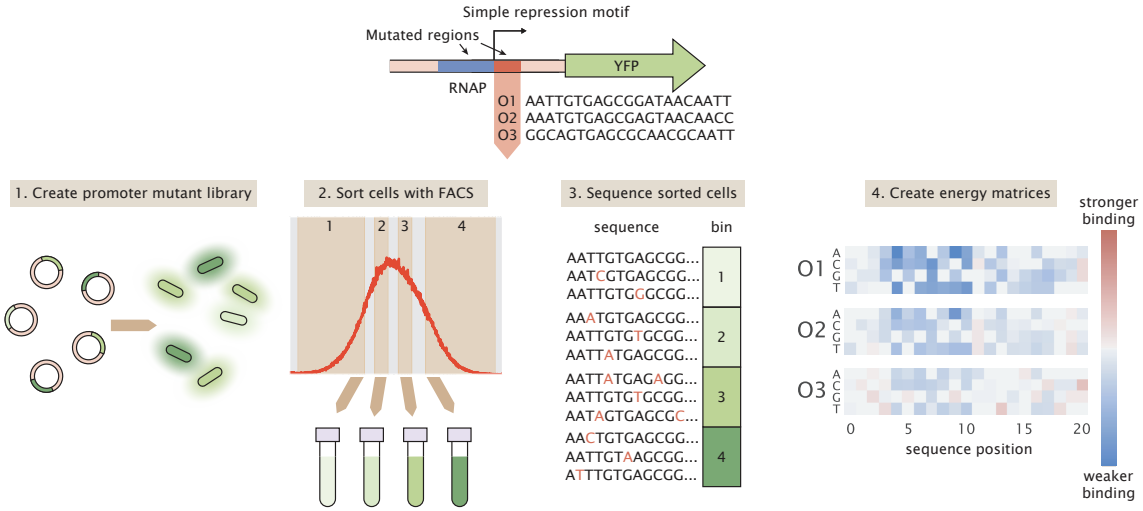

Figure 4.1: **Process flow for using Sort-Seq to obtain energy matrices.** To begin, we design a simple repression motif in which a repressor binding site is placed immediately downstream of the RNAP site. When RNAP binds, it initiates transcription of the GFP reporter gene. We analyze simple repression constructs using each of the three natural *lac* operators, O1, O2, and O3. Sort-Seq then proceeds according to the following process flow. 1. We create a mutant library in which the RNAP and operator sequences are randomly mutated at a rate of approximately 10%, and transform this library into a cell population such that each cell contains a different mutant operator sequence. 2. To measure gene expression, we sort the cell population into bins based on fluorescence level. 3. We then sequence variant promoter sequences within each bin. The bin in which each promoter is found serves as a measure of that promoter's activity. 4. From this information, we can infer an energy matrix for the repressor binding site indicating which mutations result in a higher or lower binding energy relative to the reference sequence.

Choice of reference sequence can alter the repressor's apparent sequence specificity

One might assume that affinity experiments should reveal the same binding specificity regardless of the set of binding site mutants used in the experiment. To test this possibility, we generated energy matrices using three different reference sequences. A reference sequence refers to the sequence which serves as the "wild-type" for each experiment. For each library, the promoter is mutated relative to its reference sequence. Additionally, when assigning binding energies to an energy matrix, all binding energies are calculated relative to the reference sequence. For our reference sequences we use the three natural *E. coli lac* operators (O1 = AATTGTGAGCGGATAACAATT, O2 = AAATGTGAGCGAGTAACAACC, and O3 = GGCAGTGAGCGCAACGCAATT). For our primary analysis we use energy matrix models. These models assume that each nucleotide position within a binding site contributes independently to the binding energy (see Supplemental Section 4.8 for predictions using higher-order models). Each operator has a distinct LacI binding energy, with O1 being the strongest at -15.3 k_BT, O2 being the second strongest at -13.9 k_BT, and O3 being the weakest at -9.7 k_BT [29]. The operator sequences are rather dissimilar to each other, with O2 having 5 mutations relative to O1 and O3 having 8 mutations relative to O1 (and 11 mutations relative to O2). For each library, the average operator sequence has only 2 mutations relative to the reference sequence. As a result, a library generated with O1 as the reference sequence is unlikely to share any mutant sequences with a library generated with O2 or O3 as the reference sequence. Here we assess whether dissimilar mutant libraries generated from different reference sequences produce similar energy matrices and sequence logos from their respective Sort-Seq data sets.

As shown in Figure 4.2(A), the three operators each produce qualitatively similar energy matrices, with the left side of the binding site showing greater sequence dependence than the right side, as evidenced by the larger magnitude of the binding energies assigned to each matrix position. Note that we set the binding energy of the reference sequence to 0 k_BT for these energy matrices, so that the binding energies assigned to each possible mutation are calculated relative to the reference sequence. For all energy matrices, positions 4-10 show the greatest sequence preference. This preference is reflected in the natural *lac* operator sequences themselves, as the bases from 4-10 are conserved in each of the operators. Notably, the majority of mutations available to O1 incur a penalty to binding energy, while many of the mutations available to O3 enhance the binding energy. This is consistent with the

observation that O1 has a strong binding energy while O3 has a weak binding energy.

When the energy matrices are used to produce sequence logos (see Ref. [30] and Supplemental Section 3.7), we see a consistent preference for a slightly asymmetric binding site, reflecting the fact that LacI is known to bind asymmetrically to its operators. Additionally, clear differences arise for the different operators (see Figure 4.2(B)). One of the most striking differences is the information content of each sequence logo; as the binding energy of the reference sequence grows weaker, the average information content of each nucleotide position grows smaller. Additionally, while the sequence logos derived from O1 and O2 indicate very similar sequence preferences, the preferred sequence suggested by the O3 sequence logo differs in some prominent positions. In Supplemental Section 4.9 we note that weaker binding sites exhibit a greater variation in the quality of their sequence logos; thus it may be that the O3 binding site is simply too weak to provide an informative sequence logo.

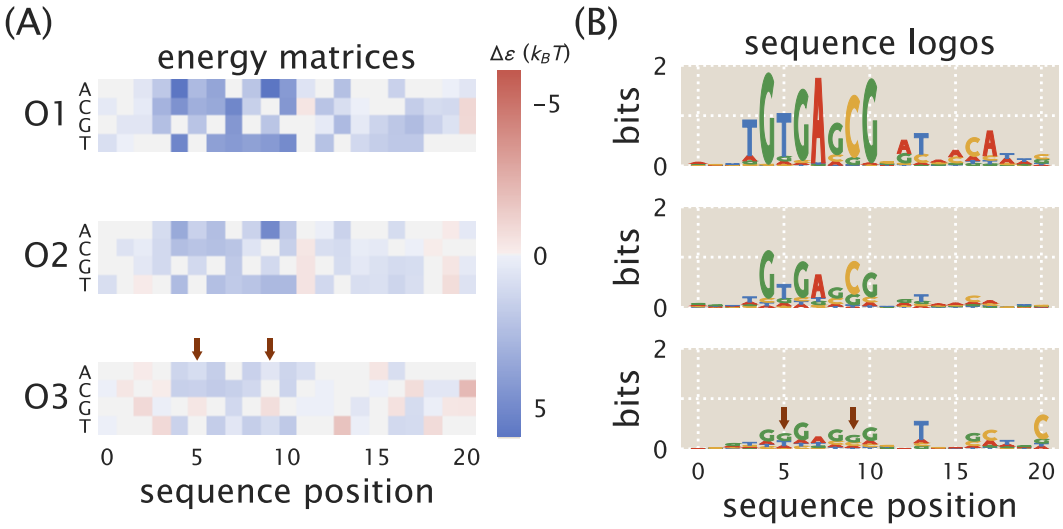

Figure 4.2: **Energy matrices and sequence logos for the natural *lac* operators.** (A) Energy matrices show how mutations can be expected to affect binding energy. Reference sequences for each energy matrix (either the O1, O2, or O3 sequence) have been set at 0 k_BT (gray squares), and the energy values at all other positions of the matrix are thus relative to the reference sequence. Red squares represent mutations that create a stronger binding energy than the reference sequence, and blue squares represent mutations that create a weaker binding energy. In columns where multiple squares are gray, this indicates that there is no significant change in binding energy relative to the reference sequence. Positions where preferred bases differ from the O1 matrix are noted with arrows. (B) While the energy matrices are qualitatively similar for all three operators, the sequence logos indicate clear differences in the information that can be provided by each operator. The O1 and O2 operators produce similar sequence logos, but the O3 sequence logo incorrectly predicts the preferred binding sequence for LacI. The O3 sequence logo also indicates a much lower information content than for O1 and O2. Positions where preferred bases differ from the O1 sequence logo are noted with arrows.

Linear energy matrix models predict measured energy values

The energy matrices obtained via Sort-Seq should allow us to map sequence to phenotype. The relevant phenotype for simple repression constructs is the degree to which the system is repressed, which can be measured using the fold-change. We define fold-change as the ratio of expression in a repressed system to expression in a system with no repressors, as described by the equation

$$\text{fold-change} = \frac{\text{expression}(R)}{\text{expression}(R = 0)}. \tag{4.1}$$

As discussed in further detail elsewhere [28, 29], the fold-change can also be computed using a thermodynamic model given by

$$\text{fold-change} = \frac{1}{1 + \frac{2R}{N_{NS}} e^{-\beta \Delta\varepsilon_R}}, \tag{4.2}$$

where R is the repressor copy number, N_{NS} is the number of nonspecific binding sites available in the genome ($\sim 4.6 \times 10^6$ in *E. coli*), and $\Delta\varepsilon_R$ is the operator binding energy. We note that this model makes the simplifying assumption that the RNAP binds weakly to the promoter.

In principle, the linear energy matrix models shown in Figure 4.2 can be used to predict the binding energy of an operator mutant. To explore the ability of energy matrices to predict the effects of mutations on operator binding strength, we designed a number of mutant operators with 1, 2, or 3 mutations relative to the O1 operator. Experimentally-determined values for the binding energies of these mutants could then be compared against values predicted by our LacI energy matrices.

To obtain experimental values for mutant binding energies, we start with chromosomally-integrated simple repression constructs for each mutant that were incorporated into strains with LacI tetramer copy numbers of $R = 11 \pm 1$, 30 ± 10, 62 ± 15, 130 ± 20, 610 ± 80, and 870 ± 170, where the error denotes the standard deviation of at least three Western blot replicates as measured in Ref. [29]. We determined the fold-change by measuring the GFP fluorescence levels of each strain by flow cytometry and substituting them into Equation 4.1. We determine each mutant's binding energy, $\Delta\varepsilon_R$, by performing a single-parameter fit of Equation 4.2 to the resulting data via nonlinear regression. Figure 4.3(A) shows several fold-change values for 1 bp, 2 bp, and 3 bp mutants overlaid with these fitted curves. To provide a sense of scale for how

inaccuracies in binding energy predictions might affect the expected fold-change, the fitted curves are surrounded by a colored region representing $\Delta\varepsilon_R \pm 1\ k_BT$.

The energy matrices derived from Sort-Seq can be used to predict the value of $\Delta\varepsilon_R$ associated with a given operator mutant, as discussed in detail in Supplemental Section 4.6. Figure 4.3(B) shows how binding energy values measured by fitting to repressor titration data compare to values predicted using energy matrices. For single base pair mutations most predictions perform well and are accurate to within 1 k_BT, with many predictions differing from the measured values by less than 0.5 k_BT. Predictions are less accurate for 2 bp or 3 bp mutations, although the majority of these predictions are still within 1.5 k_BT of the measured value.

The quality of matrix predictions degrades as mutants deviate farther from the wild-type sequence used to generate the energy matrix. To evaluate predictions for a broader range of deviations from the energy matrix, we made predictions from both the O1 energy matrix and the energy matrix for O2, which has five mutations relative to O1. This allowed us to access predictions for operators that are mutated by several base pairs relative to the matrix. In Figure 4.3(C) we show how prediction error, defined as the discrepancy in k_BT between a predicted and measured energy value, varies depending on the number of mutations relative to the wild-type binding site sequence. We find that predictions remain relatively accurate for mutants that differ by up to 4 bp relative to the wild-type sequence, with median deviations of $\sim 1.5\ k_BT$ or less from the measured binding energy. Other studies have noted that linear energy matrix models fail to accurately predict binding energies for mutants with multiple mutations relative to the reference sequence [31, 32]. Thus we find that the relatively low errors depicted in Figure 4.3(C) exceed expectations for what a linear model can achieve.

We note that energy matrix quality, as measured by the accuracy of its predictions, may be affected by the experimental design. In Supplemental Section 4.9, we assess whether energy matrix quality is affected by the LacI copy number of the background strain, and find that it has little effect on matrix quality. Additionally, we compare predictions made from energy matrices with different reference sequences (i.e. O1, O2, or O3), and find that using O1 as a reference sequence produces the most accurate energy matrices, while using O3 produces energy matrices that are almost entirely non-predictive. In Supplemental Section 4.10, we consider whether better energy matrices are made using libraries in which the entire promoter is mutated or only the operator is mutated. We find that mutating the operator alone can provide

more accurate energy matrices, though one must fit to binding energy measurements in order to convert these matrices into k_BT units.

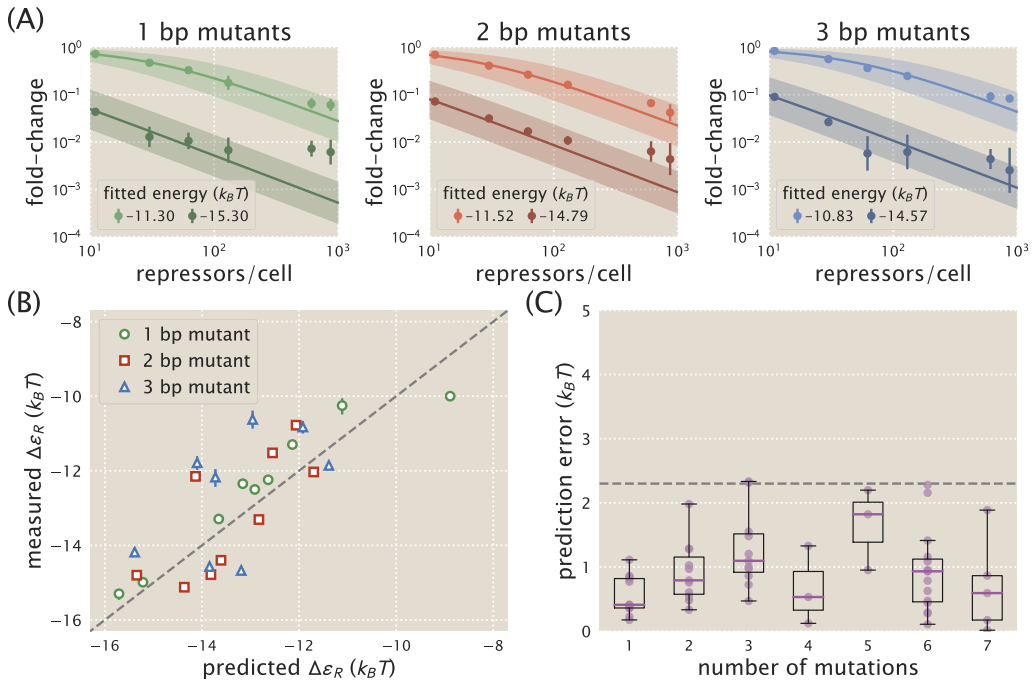

Figure 4.3: **Energy matrix predictions compared to binding energies derived from fold-change data.** (A) Fold-change data were obtained by flow cytometry for each of the mutant operators by measuring their respective fluorescence levels at multiple LacI copy numbers and normalizing by the fluorescence when $R = 0$. The solid lines in each plot represent a fold-change curve that has been fitted to the data set to obtain a binding energy measurement. The colored region surrounding each fold-change curve indicates the error in fold-change prediction that would result from an error in binding energy prediction of $\pm 1\ k_BT$. Each plot shows data and fits for two operator mutants, one weak and one strong, for 1 bp (left), 2 bp (middle), and 3 bp (right) mutants. All remaining data is shown in Supplemental Section 4.11. Approximately 30 operator mutants were measured in total. We note that expression measurements become less accurate as they grow weaker, due to autofluorescence and limitations in the flow cytometer's ability to measure weak signals. This adversely affects the accuracy of fold-change values for strongly repressed strains. (B) The measured binding energy values $\Delta\varepsilon_R$ (y axis) are plotted against binding energy values predicted from an energy matrix derived from the O1 operator (x axis). While the quality of the binding energy predictions does appear to degrade as the number of mutations relative to O1 is increased, the O1 energy matrix is still able to approximately predict the measured values. (C) Binding energies for each mutant were predicted using both the O1 and O2 energy matrices and compared against measured binding energy values. The prediction error, defined as the magnitude of the difference in k_BT between a predicted binding energy and the corresponding measured binding energy, is plotted here against the number of mutations relative to the reference sequence whose energy matrix was used to make the prediction. Each data point is shown in purple, and box plots representing the data are overlaid to clearly show the median error and variability in error. For sequences with 4 or fewer mutations, the median prediction error is consistently lower than 1.5 k_BT. The dashed horizontal line represents the point at which the error corresponds to an approximately 10-fold difference in fold-change.

Designed induction responses

Our predictive energy matrices suggest a promising strategy for addressing the challenge of genetic circuit design, which has typically relied on trial and error to achieve specific outputs [33, 34]. By contrast, previous studies have shown how thermodynamic models can be used to predict gene outputs given a set of inputs [28, 29], which can suggest appropriate inputs to produce a desired output. For example, the key inputs for the fold-change Equation 4.2 are repressor copy number R and repressor-operator binding energy $\Delta\varepsilon_R$, and one can easily use Equation 4.2 to determine a set of R and $\Delta\varepsilon_R$ values that can be used to target a desired fold-change response. Energy matrix predictions can be used to design operator sequences with a particular value of $\Delta\varepsilon_R$, thereby making it possible to tune genetic circuits and target specific phenotypes. As shown in Figure 4.3B, mutating an operator by as little as one base pair can provide a broad range of $\Delta\varepsilon_R$ values that can be predicted accurately.

One particularly useful class of simple genetic circuit, which can be layered with other genetic components to create complex logic [35], is inducible simple repression [36–39]. In such a system, an allosteric repressor can switch between an active form, which binds to an operator with high affinity, and an inactive form, which has a low affinity to the operator. An inducer may bind to the repressor and stabilize the repressor's inactive form, thereby reducing the probability that the repressor will bind to the operator and increasing the probability that RNAP will bind and initiate transcription. The result is that an inducible system can access a broad range of fold-change values simply by tuning the concentration of inducer. As discussed in Chapter 2, the fold-change of an inducible simple repression circuit can be described by the equation

$$(c) = \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta\Delta\varepsilon_{AI}}\left(1 + \frac{c}{K_I}\right)^n} \frac{2R}{N_{NS}} e^{-\beta\Delta\varepsilon_R}\right)^{-1}, \tag{4.3}$$

where c is the concentration of inducer, n is the number of inducer binding sites on the repressor, K_A and K_I are the dissociation constants of the inducer and repressor when the repressor is in its active or inactive state, respectively, and $\Delta\varepsilon_{AI}$ is the difference in free energy between the repressor's active and inactive states. In Chapter 2 we determined that these values are $K_A = 139^{+29}_{-22}$ μM, $K_I = 0.53^{+0.04}_{-0.04}$ μM, and $\Delta\varepsilon_{AI} = 4.5$ k_BT for *lac* repressor with the inducer IPTG. Where noted, superscripts and subscripts indicate the upper and lower bounds for the 95th percentile of the

parameter value distributions. There are $n = 2$ inducer binding sites on each LacI dimer.

We can use these parameter values for the *lac*-based system considered here to explore how tuning the operator-repressor binding energy $\Delta\varepsilon_R$ can alter the induction response when an effector (i.e. IPTG) is introduced to the system. Importantly, our sequence-energy mapping provides a straightforward avenue for tuning $\Delta\varepsilon_R$ by altering the binding sequence rather than mutating the repressor itself, which is much more difficult to characterize. We note that an induction response can be described by a number of key phenotypic parameters. The leakiness is the minimum fold-change when no inducer is present, given by $(c \rightarrow 0)$. The saturation is the maximum fold-change when inducer is present at saturating concentrations, given by $(c \rightarrow \infty)$. The dynamic range is the difference between the saturation and leakiness, and represents the magnitude of the induction response. The $[EC_{50}]$ is the inducer concentration at which the fold-change is equal to the midpoint of the induction response. Full expressions for these parameters are first listed in Chapter 2 and reproduced for convenience in Supplemental Section 4.12, Equations 4.13, 4.14, 4.16, and 4.17. Figures 4.4(A) and 4.4(B) show how these phenotypic parameters vary with $\Delta\varepsilon_R$ given the values of K_A, K_I, and $\Delta\varepsilon_{AI}$ listed above and the repressor copy number $R = 130$. We can see that there are inherent trade-offs between phenotypic parameter values. For instance, in this particular system one cannot tune $\Delta\varepsilon_R$ to obtain a small dynamic range (e.g. a dynamic range of 0.1) while also having an intermediate leakiness value (e.g. a leakiness of 0.4). Rather, one must design an induction response by choosing from the available phenotypes, or else alter the system by tuning additional parameters such as K_A and K_I, which requires mutating the protein itself or using a different transcription factor altogether as in Ref. [33].

To show how energy matrices can be used to design specific induction responses, we used the phenotypic trade-offs shown in Figures 4.4(A) and 4.4(B) to choose values of $\Delta\varepsilon_R$ that would provide distinct outputs. A strong binding energy lies below $\Delta\varepsilon_R \approx -14\ k_BT$, which provides a minimal leakiness level but not full saturation, and gives a high $[EC_{50}]$ value. A moderate binding energy lies in the range $\Delta\varepsilon_R \approx -14$ to $-12\ k_BT$, maximizing dynamic range and giving an intermediate $[EC_{50}]$ value. Finally, weak binding energies lie above $\Delta\varepsilon_R \approx -12\ k_BT$, which provides a narrower dynamic range and a lower $[EC_{50}]$ value. We chose six of our single base-pair mutants with predicted binding energies in these ranges.

Induction responses for each of these mutants were measured by growing cultures in the presence of varying IPTG concentrations and measuring the fold-change at each concentration, following the procedure described in Chapter 2. Figure 4.4(C-H) shows how the induction data compare against theory curves plotted using $\Delta\varepsilon_R$ values predicted from the energy matrix. For operators with stronger binding energies, the data match well with the theory curves plotted using predicted binding energies (Figure 4.4(C-E)). For operators with weaker binding energies, however, we find that the data do not match as well with the predicted theory curves (Figure 4.4(F-H)). Theory curves plotted using the measured binding energy (rather than the predicted binding energy) match well with the data, indicating that the mis-match between the data and the predicted theory curve is due to error in the predicted binding energy.

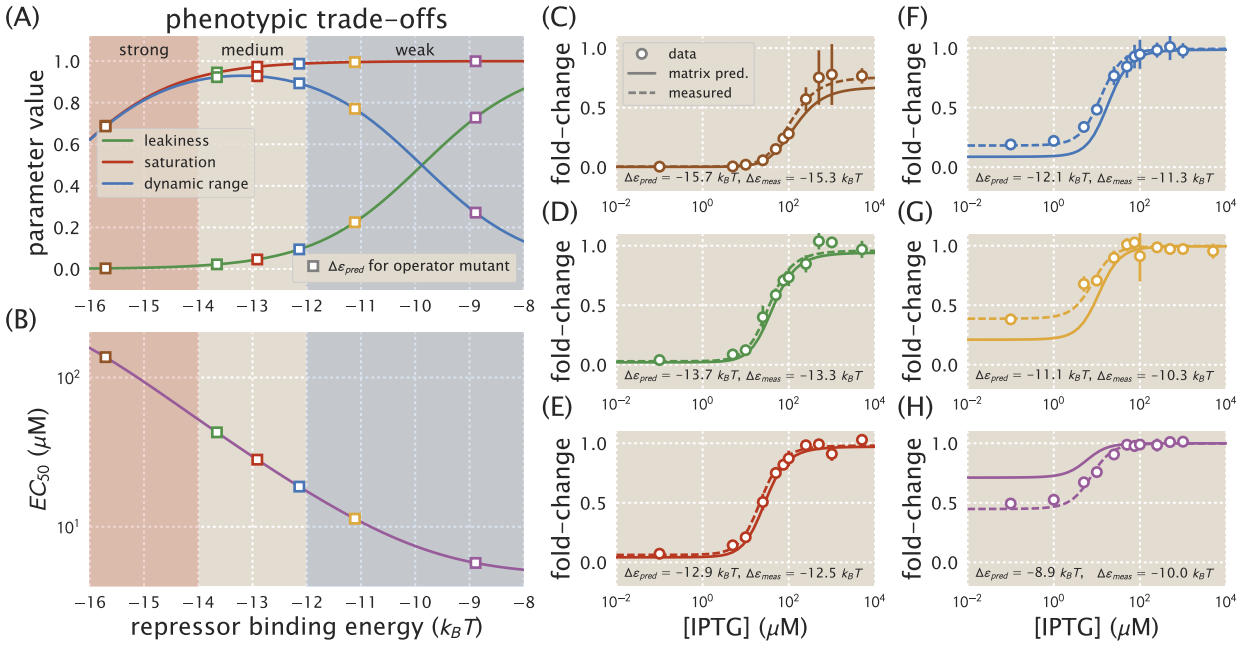

Figure 4.4: **Energy matrix predictions can be used to design phenotypic responses** Phenotypic parameters exhibit trade-offs as $\Delta\varepsilon_R$ is varied. (A) The values of the leakiness, saturation, and dynamic range are plotted as a function of transcription factor binding energy, $\Delta\varepsilon_R$, for a strain with $R = 130$. Different values of $\Delta\varepsilon_R$ fall into different binding regimes (strong, medium, or weak) with different phenotypic properties. Several operators were chosen whose predicted binding energies (squares) fall into these different binding regimes. (B) The value of the $[EC_{50}]$ is plotted as a function of $\Delta\varepsilon_R$ for a strain with $R = 130$. The $[EC_{50}]$ decreases as the value of $\Delta\varepsilon_R$ increases. (C-H) Operators with different values of $\Delta\varepsilon_R$ were chosen to have varying induction responses based on the phenotypic trade-offs shown in (A) and (B). The fold-change is shown for each operator as IPTG concentrations are varied. (C-E) For operators with stronger binding energies, the data match well with both the predicted theory curves and the theory curves based on measured binding energies. (F-H) For operators with weaker binding energies, the data match well with theory curves based on measured binding energies, but do not match as well with predicted theory curves, due to inaccuracies in the energy matrix predictions. For each of these operators, the predicted binding energy $\Delta\varepsilon_{pred}$ differs from the measured binding energy $\Delta\varepsilon_{meas}$ by $\sim 1\ k_BT$.

Analysis of amino acid-nucleotide interactions

Predictive energy matrices offer a simple way of analyzing direct interactions between amino acids and nucleotides. Mutating individual amino acids in the repressor's DNA-binding domain and then observing changes in the energy matrix makes it possible to determine how changing the amino acid composition of the DNA-binding domain alters sequence preference. If sequence specificity is altered only for specific base pairs when an amino acid is mutated, this may indicate that the amino acid interacts directly with those base pairs. While it is possible to obtain such information using binding assays [40] or labor-intensive structural biology approaches, Sort-Seq makes it possible to efficiently sample a full array of operator mutations in a single experiment. To analyze the effects of mutations on sequence specificity, we chose mutations which had previously been found to alter LacI-DNA binding properties without entirely disrupting the repressor's ability to bind DNA [40, 41]. We performed Sort-Seq using strains containing one of three LacI mutants, Y20I, Q21A, or Q21M, where the first letter indicates the wild-type amino acid, the number indicates the amino acid position, and the last letter indicates the identity of the mutated amino acid.

The energy matrices for each LacI mutant are shown in Figure 4.5(A), along with the wild-type energy matrix for comparison. Sequence logos derived from each energy matrix are shown in Figure 4.5(B). The energy matrices remain remarkably similar to one another. As with the wild-type repressor, for each of the mutant repressors we find that the left half-site of the sequence logo has a stronger sequence preference. For both Y20I and Q21M, the same sequence is preferred in the left half-site as for the wild-type LacI. This contrasts with the results from Ref. [40], in which it was found that Y20I prefers an adenine at sequence position 6, rather than the guanine preferred at this position by the wild-type repressor. As in Ref. [40], we find that an adenine is preferred at sequence position 6 for the Q21A mutant. Additionally, when comparing the left and right half-sites of each energy matrix, we find that for each mutant the preferred sequence is not entirely symmetric. Thus we see that the *lac* repressor's notable preference for a pseudo-symmetric operator is preserved in each of the mutants we tested.

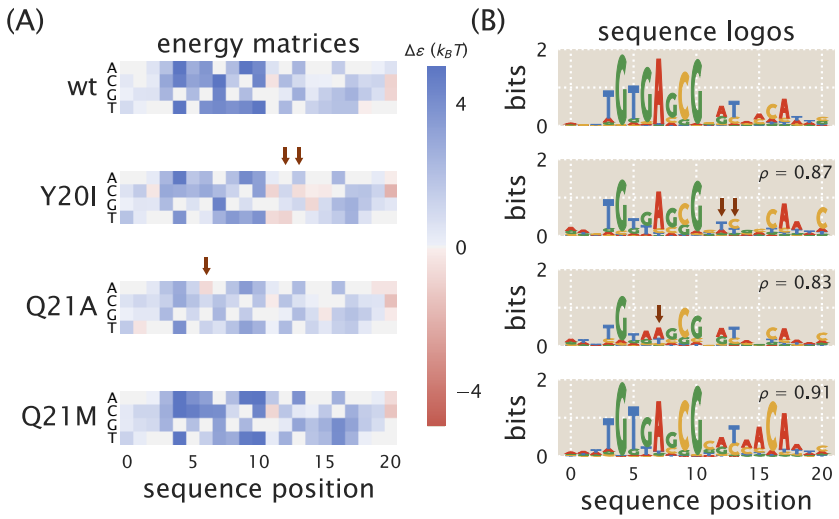

Figure 4.5: **Mutations to LacI DNA-binding domain cause subtle changes to sequence specificity.** Mutations were made to residues 20 and 21 of LacI, both of which lie within the DNA-binding domain. The mutations Y20I and Q21A weaken the repressor-operator binding energy, while the mutation Q21M strengthens the binding energy [41]. The sequence preferences of each mutant are represented as (A) energy matrices and (B) sequence logos. Y20I exhibits minor changes to specificity in low-information regions of the binding site, and Q21A experiences a change to specificity within a high-information region of the binding site (see arrows). Specifically, Q21A prefers A at operator position 6 while the wild-type repressor prefers G at this position. The Pearson's corellation coefficient ρ is noted for each mutant, calculated by comparing the energy matrix values for each mutant to the wild-type energy matrix values. For comparison, replicates of the O1 energy matrix with wild-type LacI all have values of $\rho \geq 0.93$ relative to one another (see Supplemental Section 4.9).

Binding site context can influence a transcription factor's binding specificity

In this work we have used the *lac* system to demonstrate how Sort-Seq can be used to map binding site sequence to binding energy, and we used these mappings to rationally design novel genetic circuit elements and identify the effects of amino acid mutations on LacI's sequence specificity. Importantly, this approach is not specific to the *lac* system and can be applied to any system in which transcription factors alter gene expression by binding to DNA within the promoter region. In Chapter 3 we showed how Sort-Seq could be used alongside mass spectrometry to determine the locations of transcription factor binding sites in a promoter of interest and identify which transcription factors bind to these sites. We generated energy matrices for a number of transcription factors (e.g. RelBE, MarA, PurR, XylR, and others), but we did not use these energy matrices to perform quantitative analyses as we do here. Here we analyze selected energy matrices from Chapter 3 to show how energy matrices can be used to understand transcriptional activity in promoters with varied architectures beyond simple repression.

One of the questions we wish to answer is to what extent altering the context of a binding site within a regulatory architecture will alter sequence specificity. One hypothesis is that a transcription factor's preferred binding sequence will remain the same regardless of how its binding site is positioned within the regulatory architecture. However, it is known that factors beyond the core operator sequence, such as flanking sequences and DNA shape, can affect sequence specificity [19, 42, 43]. Additionally, interactions with other proteins may alter the way a transcription factor contacts the DNA, which could affect sequence specificity as well [44]. It is important to know whether a transcription factor's specificity is sensitive to the context of the binding site within the promoter architecture, as this determines the extent to which an energy matrix can be used to analyze binding sites throughout the genome. Additionally, observing how sequence specificities change with binding site context may alert us to changes in regulatory mechanisms as the operator is moved to different positions in the promoter.

In Chapter 3, we used Sort-Seq to obtain energy matrices and sequence logos for the transcription factors XylR and PurR in the context of the natural promoters for *xylE* and *purT*, respectively. The *xylE* promoter has two XylR binding sites directly adjacent to one another, allowing us to compare these two energy matrices against each other. In this context, we find that XylR appears to act as an activator in tandem with a CRP binding site. Sequence logos for the two XylR binding sites are shown

in Figure 4.6A. The energy matrices and sequence logos for these binding sites have some significant dissimilarities. Dissimilarities are particularly notable at positions 6-8, where the left-hand site prefers "TTT" and the right-hand site prefers "AAA". In the *xylE* promoter the left-hand XylR site is adjacent to a CRP site, while the right-hand XylR site is adjacent to the RNAP site. The close proximity of these binding sites suggests that there may be direct interactions between proteins, which could alter how each XylR interacts with the DNA, thus altering sequence preferences. The Pearson's corellation coefficient ρ between the two energy matrices is $\rho = 0.57$.

In Chapter 3 we find that PurR acts as a repressor in the *purT* promoter, with a single binding site between the -10 and -35 sites. In order to compare the associated energy matrix with a PurR energy matrix from a different regulatory context, here we create a synthetic promoter in which the PurR binding site has been moved directly downstream of the RNAP site. This should continue to be a simple repression architecture in which repressor binding occludes RNAP binding, but the change in operator position may alter the repressor's interaction with the DNA. Sequence logos for both PurR binding sites are shown in Figure 4.6B. The two PurR sequence logos are very similar to one another, indicating no significant changes in the interactions between the repressor and the DNA. We calculate the Pearson's corellation coefficient between the two energy matrices to be $\rho = 0.90$, which is significantly higher than the value calculated for the two XylR energy matrices.

We additionally performed Sort-Seq on a LacI simple repression construct in which the *lac* operator was placed upstream of the RNAP binding site rather than downstream. In Ref. [26] it is shown that LacI binding to an upstream operator still represses, but whereas a downstream operator represses by preventing RNAP from binding, an upstream operator appears to directly contact a bound RNAP and prevent it from escaping the promoter. Moreover, an upstream operator's binding strength does not directly correspond with the level of repression associated with the promoter. These factors make repression by an upstream *lac* operator an interesting architecture to compare with repression by a downstream *lac* operator. Sequence logos for the upstream and downstream LacI binding sites are shown in Figure 4.6(C). These logos are very similar to one another, despite the fact that the repression mechanisms and protein interactions differ for these two architectures. The Pearson's corellation coefficient between the two matrices is $\rho = 0.95$.

Because a definitive thermodynamic model was not available for all of the architectures examined in Figure 4.6, the energy matrices used to make the sequence logos

were scaled using a theoretical "average" binding penalty derived from a statistical mechanical analysis of transcriptional regulation (see Supplemental Section 4.7). Supplemental Section 4.5 shows the wild-type binding sites that act as reference sequences for the sequence logos.

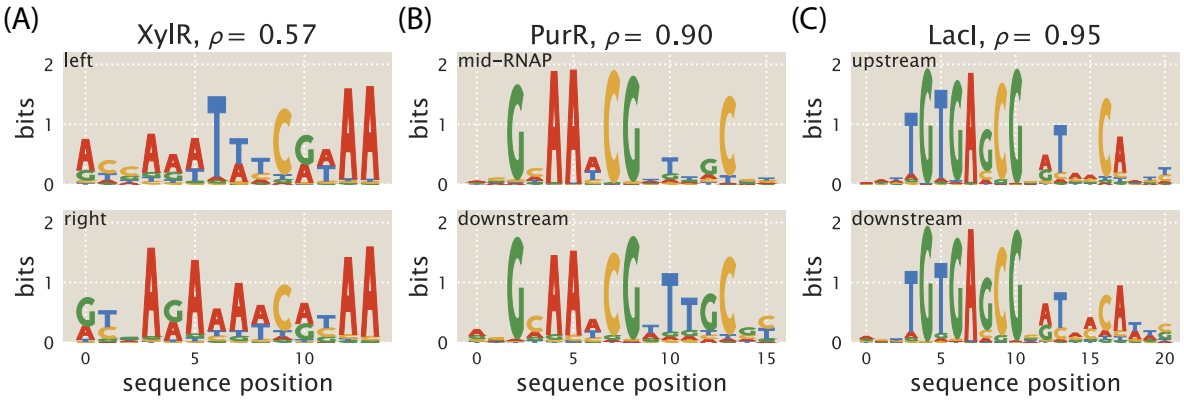

Figure 4.6: **Regulatory context can alter sequence preference.** Sequence logos were obtained for the same transcription factors in different regulatory contexts and compared against one another. The Pearson's corellation coefficient ρ between energy matrices is noted for each pair of binding sites. (A) Sequence logos are shown for the two adjacent binding sites for the activator XylR in the *xylE* promoter, shown schematically at top. The sequence logos for the two binding sites indicate that they have significantly different sequence preferences. (B) Sequence logos are shown for the PurR binding site in the *purT* promoter and a PurR binding site for a synthetic simple repression promoter in which the binding site is positioned differently, shown schematically at top. The sequence logos for the two binding sites indicate nearly identical sequence preferences. (C) Sequence logos are shown for a LacI binding site upstream of the RNAP binding site and a LacI binding site downstream of the RNAP. Although regulatory mechanisms differ between these two binding sites, their sequence logos are nearly identical.

4.3 Discussion

In this work, we apply quantitative modeling to *in vivo* experimental techniques to analyze interactions between transcription factors and their binding sites under multiple conditions. As an example of how our approach might be used to analyze a transcription factor's sequence-specific binding energy, we used Sort-Seq to create energy matrices that map DNA sequence to binding energy for the *lac* repressor (Figure 4.2). We performed this work in the context of a simple repression architecture, which is widespread among bacterial promoters [45] and is frequently used in synthetic biology [39, 46, 47]. We test our model's predictions against binding energies inferred from fold-change measurements of roughly 30 *lac* operator mutants (Figure 4.3). These predictions proved to be approximately accurate, even for operators with multiple mutations.

Because we are able to accurately predict operator binding energies, our sequence-energy mappings can be used to design specific regulatory responses, which is of great utility to synthetic biology. We combine energy matrices with the thermodynamic model of inducible simple repression introduced in Chapter 2 to design induction curves, as demonstrated in Figure 4.4. We note that in spite of the overall success of our predictions, there remain some predictions that are significantly different from the measured values (see the outliers in Figure 4.3(C)). Such inaccuracies are particularly problematic when using energy matrices for design applications, as discrepancies between a system's expected and actual response may render a designed system unsuitable for its intended application. We can see examples of this in Figure 4.4(F-H). The prediction curves corresponding to operators with weaker binding energies do not accurately describe the data, with the data exhibiting higher or lower leakiness values than was predicted. If the leakiness is a vital parameter in the designed system, then such a mis-match could cause the system to fail.

We also explore how sequence specificity is altered when transcription factor amino acids are mutated. To do this, we repeat our Sort-Seq experiments in bacterial strains expressing LacI mutants in which the DNA-binding domain has been altered (Figure 4.5). Because all nucleotides in the binding site are mutated with some frequency in Sort-Seq experiments, we are able to identify changes in specificity throughout the entire binding site. Other methods for analyzing the sequence preference of transcription factor mutants tend to be more laborious and less fine-grained, often focusing on a small set of nucleotides within the binding site. These include binding experiments between DNA mutants and protein mutants [40], gene expres-

sion experiments using chimeric transcription factor proteins [48], and comparative genomics [49].

We further explore how regulatory context alters sequence specificity. We generate sequence logos from energy matrices obtained for the transcription factors XylR, PurR, and LacI in different regulatory contexts, as shown in Figure 4.6. We find that the two adjacent XylR binding sites exhibit significantly different binding specificities, while the simple repression constructs analyzed for PurR and LacI have nearly identical sequence specificities. By itself, our method is unable to determine the causes of context-dependent changes in sequence specificity, though it is known that DNA shape or binding to cofactors can alter a transcription factor's specificity [42–44]. Rather, our approach can be used to determine whether a given binding site's sequence preferences diverge from the "standard" sequence specificity for the relevant transcription factor, and further experiments (such as SELEX-seq in the presence of a transcription factor and possible cofactors [44]) can be performed to determine the cause of the change in sequence specificity.

A major advantage of our *in vivo* approach is that it allows us to analyze transcription factors in their natural context, in the presence of interacting proteins, small molecules, and DNA shape effects. This is especially important when analyzing regulatory regions that have not been previously annotated, as was the case for the XylR and PurR matrices obtained in Chapter 3. However, a clear advantage of *in vitro* approaches is that they can accurately measure low-affinity binding sites [12, 13, 15]. When using our *in vivo* approach, weaker reference sequences produce energy matrices with variable quality and are more likely to make poor predictions (see Supplemental Section 4.9). However, accuracy may be improved by investigating ways to reduce the experimental noise associated with *in vivo* systems, for instance by incorporating promoter constructs as single copies in the chromosome rather than multiple copies on plasmid, for example using the "landing pad" technique described in Ref. [50].

This work provides a foundation for further studies that would benefit from sequence-energy mappings. For example, our analysis of three LacI amino acid mutants could be expanded to include a full array of DNA-binding mutants, which would allow one to make inferences regarding repressor-operator coevolution. Additionally, while we make extensive use of LacI in the present work, similar analyses could be performed with any transcription factor, making it possible to improve upon the genomically-inferred sequence logos presently available for many transcription factors. Further,

for cases in which it is known that sequence specificity is affected by DNA shape, flanking sequences, cofactor binding, or other factors outside of the operator binding sequence, our approach can be used to obtain a finely-detailed map of the effects on sequence specificity. Finally, we note that one of the primary strengths of our approach is that it can be used to elucidate the transcriptional regulation of a gene with a previously-unknown regulatory architecture. As we have already shown in Chapter 3, Sort-Seq can be combined with mass spectrometry to identify transcription factor binding sites and those sites' regulatory roles for any gene of interest. Here we show that data sets obtained in this manner can also be used to map sequence to binding energy, thus showing that a single experiment can be used to characterize multiple aspects of a previously-unannotated regulatory sequence. Futhermore, our approach does not rely specifically on the Sort-Seq technique used here, but can be adapted to multiple experimental designs, such as RNA-seq based MPRAs that have been demonstrated in multiple model systems [7, 51–53]. Over time, we envision incorporating high-throughput synthesis and analysis techniques to adapt our approach for genome-wide studies in both prokaryotes and eukaryotes.

4.4 Methods

Sort-Seq libraries

To generate promoter libraries for Sort-Seq, mutagenized oligonucleotide pools were purchased from Integrated DNA Technologies (Coralville, IA). These consisted of single-stranded DNA containing the *lacUV5* promoter and LacI operator plus 20 bp on each end for PCR amplification and Gibson Assembly. Either both the *lacUV5* promoter and LacI binding site or only the LacI binding site was mutated with a ten percent mutation rate per nucleotide. These oligonucleotides were amplified by PCR and inserted back into the pUA66-operator-GFP construct using Gibson Assembly. To achieve high transformation efficiency, reaction buffer components from the Gibson Assembly reaction were removed by drop dialysis for 90 minutes and cells were transformed by electroporation of freshly prepared cells. Following an initial outgrowth in SOC media, cells were diluted with 50 mL LB media and grown overnight under kanamycin selection. Transformation typically yielded 10^6 – 10^7 transformants as assessed by plating 100 μL of cells diluted 1:10^4 onto an LB plate containing kanamycin and counting the resulting colonies.

DNA Constructs for fold-change measurements of mutant operators

Simple repression motifs used in fold-change measurements were adapted from those in Garcia *et al.*[29]. Briefly, a simple repression construct with the O1 operator sequence was cloned into a pZS25 plasmid background directly downstream of a *lacUV5* promoter, driving expression of a YFP gene when the operator is not bound by LacI. This plasmid contains a kanamycin resistance gene for selection. Mutant LacI operator constructs (listed in Table 4.1) were generated by PCR amplification of the *lacUV5* O1-YFP plasmid using primers containing the point mutations as well as sufficient overlap for re-circularizing the amplified DNA by one-piece Gibson Assembly.

A second construct was generated to express LacI at a specified copy number. Specifically, *lacI* was cloned into a pZS3*1 background that provides constitutive expression of LacI from a $P_{LtetO-1}$ promoter [54]. This plasmid contains a chloramphenicol resistance gene for selection. The LacI copy number is controlled by mutating the ribosomal binding site (RBS) for the *lacI* gene as described in [55] using site-directed mutagenesis (Quickchange II; Stratagene, San Diego, CA) and further detailed in [29]. Here, we mutated the RBS such that it would produce a LacI copy number of $\sim$ 130 tetramers once the construct had been integrated into the chromosome.

Once the plasmids had been generated, the promoter and *lacI* constructs were each amplified by PCR and integrated into the chromosome by lambda-red recombineering using the pSIM6 expression plasmid [56]. The promoter construct and YFP gene were inserted into the *galK* locus in the *E. coli* genome and the *lacI* construct was inserted into the *ybcN* locus.

Construction of LacI Amino Acid Mutants

As previously mentioned, wild-type *lacI* was cloned into a pZS3*1 background providing constitutive expression of LacI, with the LacI copy number mediated by a mutated RBS. We used the RBS corresponding to a LacI tetramer copy number of $\sim$ 130 for each mutant. To create DNA-binding mutants for LacI we used site-directed mutagenesis (Quickchange II; Stratagene, San Diego, CA) using the mutagenesis primers listed in Table 4.2. We mutated the amino acid Y to I at position 20 and Q to A or M at position 21. We chose these mutations based on data from previous studies [40, 41], though we note that our amino acid numbering system is shifted by +3 relative to the mutants in these previous studies since we use a slightly different version of *lacI*. As with the wild-type *lacI*, we integrate the mutants into the genome at the *ybcN* locus by lambda-red recombineering using the pSIM6 expression plasmid.

Bacterial Strains

E. coli strains used in this work were derived from K12 MG1655. To generate strains with different LacI copy number, the *lacI* constructs were integrated into a strain that additionally has the entire *lacI* and *lacZYA* operons removed from the chromosome. These constructs were integrated at the *ybcN* chromosomal location. This resulted in strains containing mean LacI tetramer copy numbers of $R = 11 \pm 2$, 30 ± 10, 62 ± 15, 130 ± 20, 610 ± 80, and 870 ± 170, where the error denotes the standard deviation of at least three replicates as measured by quantitative western blots in Ref. [29].

For Sort-Seq experiments, plasmid promoter libraries were constructed as described below and then transformed into the strains with $R = 30, 62, 130$ or 610. For fold-change measurements, each O1 operator mutant was integrated into strains containing each of the listed LacI copy numbers. These simple repression constructs were chromosomally integrated at the *galK* chromosomal location via lambda red recombineering. Generation of the final strains containing a simple repression motif and a specific LacI copy number was achieved by P1 transduction. For each LacI

titration experiment, we also generated a strain in which the operator-YFP construct had been integrated, but the *lacI* and *lacZYA* operons had been removed entirely. This provided us with a fluorescence expression measurement corresponding to $R = 0$, which is necessary for calculation of fold-change.

Sort-Seq fluorescence sorting

For each Sort-Seq experiment, cells were grown to saturation in lysogeny broth (LB) and then diluted 1:10,000 into minimal M9 + 0.5% glucose for overnight growth. Once these cultures reached an OD of 0.2-0.3 the cells were washed three times with PBS by centrifugation at 4000 rpm for 10 minutes at 4°C. They were then diluted two-fold with PBS to reach an approximate OD of 0.1-0.15. These cells were then passed through a 40 μm cell strainer to eliminate any large clumps of cells.

A Beckman Coulter MoFlo XDP cell sorter was used to obtain initial fluorescence histograms of 500,000 events per library in the FL1 fluorescence channel with a PMT voltage of 800 V and a gain of 10. The histograms were used to set four binning gates that each covered $\sim$ 15% of the histogram. 500,000 cells were collected into each of the four bins. Finally, sorted cells were regrown overnight in 10 mL of LB media, under kanamycin selection.

Sort-Seq sequencing and data analysis

Overnight cultures from each sorted bin were miniprepped (Qiagen, Germany), and PCR was used to amplify the mutated region from each plasmid for Illumina sequencing. The primers contained Illumina adapter sequences as well as barcode sequences that were unique to each fluorescence bin, enabling pooling of the sorted samples. Sequencing was performed by either the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech or NGX Bio (San Fransisco, CA). Single-end 100bp or paired-end 150bp flow cells were used, with about 500,000 non-unique sequences collected per library bin. After performing a quality check and filtering for sequences whose PHRED score was greater than 20 for each base pair, the total number of useful reads per bin was approximately 300,000 to 500,000 per million reads requested. Energy weight matrices for binding by LacI and RNAP were inferred using Bayesian parameter estimation with a error-model-averaged likelihood as previously described [27, 57] and further detailed in Supplemental Section 4.6.

Fold-change measurements by flow cytometry

Fold-change measurements were collected as previously described in Chapter 2 on a MACSquant Analyzer 10 Flow Cytometer (Miltenyi Biotec, Germany). Briefly, YFP fluorescence measurements were collected using 488nm laser excitation, with a 525/50 nm emission filter. Settings in the instrument panel for the laser were as follows: trigger on FSC (linear, 423V), SSC (linear, 537 V), and B1 laser (hlog, 790V). Before each experiment the MACSquant was calibrated using MACSQuant Calibration Beads (Miltenyi Biotec, CAT NO. 130-093-607). Cells were grown to OD 0.2-0.3 and then diluted tenfold into ice-cold minimal M9 + 0.5% glucose. Cells were then automatically sampled from a 96-well plate kept at approximately 4° - 10°C using a MACS Chill 96 Rack (Miltenyi Biotec, CAT NO. 130-094-459) at a flow rate of 2,000 - 6,000 measurements per second.

For those measurements that were taken for IPTG induction curves, cells were grown as above with the addition of an appropriate concentration of IPTG (Isopropyl β-D-1 thiogalactopyranoside Dioxane Free, Research Products International). For each IPTG concentration, a stock of 100-fold concentrated IPTG in double distilled water was prepared and partitioned into 100 μL aliquots. The same parent stock was used for all induction experiments described in this work.

The fold-change in gene expression was calculated by taking the ratio of the mean YFP expression of the population of cells in the presence of LacI to that in the absence of LacI. Since the measured fluorescence intensity of each cell also includes autofluorescence which is present even in the absence of YFP, we account for this background by computing the fold change as

$$\text{fold-change} = \frac{\langle I_{R>0} \rangle - \langle I_{\text{auto}} \rangle}{\langle I_{R=0} \rangle - \langle I_{\text{auto}} \rangle}, \tag{4.4}$$

where $\langle I_{R>0} \rangle$ is the average cell YFP intensity in the presence of repressor, $\langle I_{R=0} \rangle$ is the average cell YFP intensity in the absence of repressor, and $\langle I_{\text{auto}} \rangle$ is the average cell autofluorescence intensity as determined by measuring the fluorescence of cells in which $R = 0$ and there is no fluorescent reporter.

Data curation

All data was collected, stored, and preserved using the Git version control software in combination with off-site storage and hosting website GitHub. Raw flow cytometry data files (`.fcs` and `.csv`) files were stored on-site under redundant storage. Due to size limitations, these files are available upon request. Sequencing data is available through the NCBI website under accession number SAMN08930313.

Table 4.1: **Mutant operator sequences.** Each of the listed operator sequences were used to evaluate energy matrix predictions. They are mutated relative to the O1 *lac* operator. The predicted binding energy was generated using the matrix with an O1 reference sequence with $R = 130$ LacI tetramers in the background strain.

Sequence	**Predicted** $\Delta\varepsilon_R$ (k_BT)	**Measured** $\Delta\varepsilon_R$ (k_BT)
1 bp mutants:		
AATTGTGAGCGGAGAACAATT	-12.63	-12.24
AATTGTGAGCGCATAACAATT	-15.71	-15.30
AATTGTGAGCGGATCACAATT	-15.22	-14.99
AATTGTGAGCGGAAAACAATT	-12.91	-12.50
AATTGCGAGCGGATAACAATT	-12.14	-11.30
AATTGTGAGGGGATAACAATT	-13.16	-12.35
AATTGTGAGCGGATATCAATT	-13.66	-13.29
AATTGTGAGCAGATAACAATT	-11.11	-10.25
AATTGTGAGAGGATAACAATT	-8.89	-10.00
2 bp mutants:		
AATTGTGAGCGGGTAACAACT	-13.82	-14.79
AAATGTGAGCGGATAACAACT	-13.61	-14.40
AATTGTGAGCGAGTAACAATT	-14.36	-15.12
ATTTGTGAGCGGAGAACAATT	-12.55	-11.52
CATTGTGAGCGCATAACAATT	-15.34	-14.80
AATTGTGAGCGGAACACAATT	-12.83	-13.31
AATTGTGAGCGGAATACAATT	-11.70	-12.03
AATTGCGAGCGGATAACAAAT	-12.06	-10.78
AATTGTGAGGGGATAACAATC	-14.13	-12.15
3 bp mutants:		
AAATGTGAGCGAGTAACAATT	-13.84	-14.57
AATTGTGAGCGAGTAACAACT	-13.19	-14.67
ATTTGTGAGCGAAGAACAATT	-11.92	-10.83
CATTGTGAGCGCATAACATTT	-15.39	-14.18
AATTGTGAGCGGAACACAATG	-13.72	-12.17
AATTGTGAGCGGGATACAATT	-11.39	-11.86
AATTGCGAGCGGATAACAAAG	-12.96	-10.62
AATTGTGAGGGTATAACAATC	-14.10	-11.79

Table 4.2: **Primers used in this work.** The listed primer sequences were used to generate plasmids for Sort-Seq experiments or for use in creating strains with mutated operators or LacI.

Name	Sequence	Comments
lac_ins_fwd	CCCTTTCGTCTTCAC	Used to amplify *lac* promoter insert for Gibson
lac_ins_rev	CCTTTACTCATATGTATATCTCCTTTTAAATCTAGAGGAT	Used to amplify *lac* promoter insert for Gibson
pUA66_frameshift_fwd	GATATACATATGAGTAAAGGAGAAGAACTT	Used to amplify pUA66 vector for Gibson
pUA66_rev	TCGAGGTGAAGACGAAAG	Used to amplify pUA66 vector for Gibson
GCMWC-001_Q21_rev	CCGGCATACTCTGCGACA	Mutagenesis primer for LacI residue 21
GCMWC-002_Q21M	GTGTCTCTTATATGACCGTTTCCCGC	Mutagenesis primer for LacI residue 21 Q → M
GCMWC-003_Q21A	TGTCTCTTATGCGACCGTTTCCCGC	Mutagenesis primer for LacI residue 21 Q → A
GCMWC-009_Y20_rev	GCATACTCTGCGACATCGTATAAC	Mutagenesis primer for LacI residue 20
GCMWC-010_Y20I	CGGTGTCTCTATTCAGACCGTTTC	Mutagenesis primer for LacI residue 20 Y → I

Acknowledgements

Access to the Miltenyi Biotec MACSquant Analyzer 10 Flow Cytometer was graciously provided by the Pamela Björkman lab at Caltech. Access to the Beckman-Coulter MoFlo XDP cell sorter was provided by the Tirrell lab, with ongoing technical support from Tirrell lab members Seth Lieblich and Bradley Silverman. This work was supported by the National Institutes of Health DP1 OD000217 (Director's Pioneer Award) and 1R35 GM118043-01 (MIRA).

4.5 Supplemental Information: Sequences used in this work

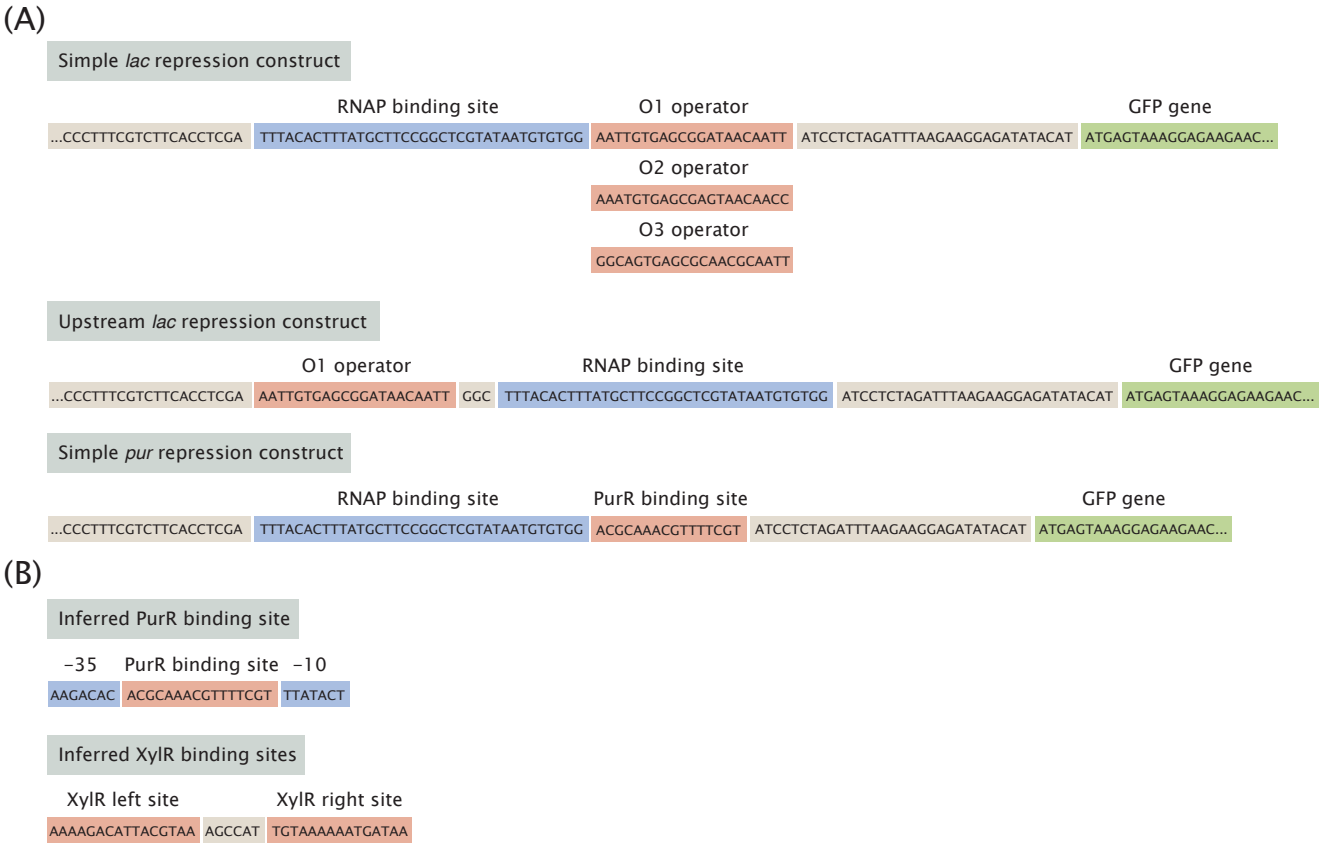

Figure 4.7: **List of wild-type reporter constructs.** (A) Wild-type versions of reporter constructs that were used either for Sort-Seq (all) or for measuring operator mutant binding energies (simple *lac* repression, though we note that constructs used for fold-change measurements used YFP rather than GFP). (B) Wild-type versions of sequences that were inferred for PurR and XylR in Chapter 3.

4.6 Supplemental Information: Bayesian Inference of Energy Matrix Models

We use Sort-Seq data to generate energy matrices that map sequence to binding energy. As discussed in Refs. [27, 58], one can infer these energy matrices by Bayesian parameter estimation using the observation that for large data sets,

$$p(\text{data} \mid \text{model}) \propto 2^{NI(\sigma;\mu)}, \tag{4.5}$$

where N is the number of data points and $I(\sigma;\mu)$ represents the mutual information between the promoter sequence σ and the fluorescence bin μ. Using a method discussed in detail in Refs. [27, 59], we use a Markov Chain Monte Carlo (MCMC) algorithm to infer a set of energy values (in arbitrary units) for each energy matrix position that maximizes the mutual information between binding site sequence and fluorescence bin. This inference is performed using the MPAthic software package [60].

In order to convert energy matrices into absolute energy units (such as the k_BT units used in this work), one must obtain a scaling factor that can be applied to the matrix. To obtain this scaling factor, we first observe that energy matrices derived from Sort-Seq can be used to predict the binding energy associated with a given operator mutant ($\Delta\varepsilon_R$) using the linear equation

$$\Delta\varepsilon_R = \alpha\varepsilon_{\text{mat}} + \Delta\varepsilon_{\text{wt}}, \tag{4.6}$$

where ε_{mat} is the energy value obtained by summing the matrix elements associated with a sequence, α is a scaling factor that converts the matrix values into k_BT units, and $\Delta\varepsilon_{\text{wt}}$ is the binding energy associated with the wild-type operator. The values of the matrix positions associated with the wild-type sequence are fixed at 0 k_BT, so that $\varepsilon_{\text{mat}} = 0$ for the wild-type sequence. Thus, $\alpha\varepsilon_{\text{mat}}$ can be interpreted as the change in binding energy relative to the wild-type caused by the specific mutations in the sequence of interest. The value of α can be determined in a number of ways (as discussed further in Supplemental Section 4.7), but the method employed in the main text is to use Bayesian parameter estimation by MCMC. The advantage of this method is that if a thermodynamic model for the promoter is known, one can use the Sort-Seq data to infer the value of α without having to perform any additional experiments. Here we describe in detail how MCMC is used to infer a value for α.

If the energy matrix is properly converted into k_BT units, then one can use energy matrix predictions, along with a thermodynamic model for gene expression, to

discern which fluorescence bin a given promoter sequence should have fallen into. We discuss above how one can infer the energy matrix parameters by maximizing the mutual information between sequence and expression bin. Similarly, we can obtain an estimate for α by finding the value of α that maximizes the mutual information between the Sort-Seq data and the expression predictions from the matrix and thermodynamic model. For the thermodynamic model, we begin with the expression for p_{bound} for a simple repression system,

$$p_{bound} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}}{1 + \frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P} + \frac{2R}{N_{NS}} e^{-\beta \Delta\varepsilon_R}}, \tag{4.7}$$

where P is the number of RNAP molecules in the system, N_{NS} is the number of nonspecific binding sites available in the system (i.e. the length of the genome), R is the number of repressors in the system, $\Delta\varepsilon_P$ is the binding energy of RNAP to its binding site, and $\Delta\varepsilon_R$ is the binding energy of the repressor to its binding site. We can rearrange this equation to make it easier to work with. First, we divide the top and bottom by the numerator, giving us

$$p_{bound} = \frac{1}{1 + \frac{1 + \frac{2R}{N_{NS}} e^{-\beta \Delta\varepsilon_R}}{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}}}. \tag{4.8}$$

Importantly, in order to evaluate the mutual information between $\Delta\varepsilon_R$ and p_{bound}, it is not necessary to adhere to the full expression for p_{bound}. Rather, we can manipulate the expression in ways that make it easier for us to work with, provided that the mutual information between $\Delta\varepsilon_R$ and p_{bound} is preserved. As noted in [57], the mutual information is preserved provided that any manipulations to the expression do not disrupt the rank ordering of an expression's values as the value of $\Delta\varepsilon_R$ is varied. We note that the term $\frac{1 + \frac{2R}{N_{NS}} e^{-\beta \Delta\varepsilon_R}}{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}}$ has the same rank ordering as the full expression for p_{bound}. Furthermore, taking the log of this term will also not affect the rank ordering, and it will make the calculation simpler, so we take the log to get an expression which we will refer to as p'_{bound}, giving us

$$p'_{bound} = \ln\left(1 + \frac{2R}{N_{NS}} e^{-\beta \Delta\varepsilon_R}\right) - \ln\left(\frac{P}{N_{NS}}\right) + \beta \Delta\varepsilon_P. \tag{4.9}$$

We observe that the constant $\ln\left(\frac{P}{N_{NS}}\right)$ also does not affect rank ordering, so we can drop this term. Additionally, we recall that $\Delta\varepsilon_R = \alpha \epsilon_{mat,R} + \Delta\varepsilon_{wt,R}$. Likewise, we

can say that $\Delta\varepsilon_P = \gamma\epsilon_{mat,P} + \Delta\varepsilon_{wt,P}$, where γ is the scaling factor for the RNAP matrix. As before, we can drop the constant $\Delta\varepsilon_{wt,P}$ as it will not affect rank ordering. This leaves us with the expression

$$p'_{bound} = \ln\left(1 + \frac{2R}{N_{NS}} e^{-\beta(\alpha\varepsilon_{mat,R} - \Delta\varepsilon_{wt,R})}\right) + \beta\gamma\varepsilon_{mat,P}. \tag{4.10}$$

With this expression in hand we can sample values of γ and α to identify values that maximize the mutual information between p'_{bound} and the expression bin which a particular sequence was sorted into during Sort-Seq. Note that while the rest of the discussion will focus on α, a value for γ comes out of this analysis as well.

The mutual information surface is very rough, with many peaks, so we need to use a method which can avoid getting stuck in local maxima. We use a parallel tempering MCMC algorithm (PTMC) to achieve this. The parallel tempering MCMC algorithm works by randomly sampling possible values for α and rejecting the value with some probability if it does not increase the mutual information relative to the previous sampled value of α. In this respect it is similar to a "standard" MCMC algorithm. By contrast with a standard MCMC algorithm, a parallel tempering algorithm runs multiple chains at once at different temperatures. In our case, we use 10 different temperatures ranging from $\beta = 0.02$ to $\beta = 4$ on a log scale, where $\beta = 1/k_BT$. Periodically throughout the MCMC run, the current α values from different temperature chains will swap. This allows the algorithm to sample α values at different levels of precision. Specifically, the high temperature chains will explore widely and not get stuck in local minima, while the low temperature chains will then carefully explore the peak that was found by the high temperature chain. The output is a distribution of values, and we take the median of this distribution to obtain our estimate for α.

4.7 Supplemental Information: Alternate Methods for Obtaining Energy Matrix Scaling Factor

As discussed in Supplemental Section 4.6, in order to convert an energy matrix into k_BT units one must infer an appropriate scaling factor α. In the main text we primarily use Bayesian parameter estimation by MCMC to infer this factor, but other methods can be used as well. Here we discuss two alternative methods: least squares regression to measured binding energy values, and calibrating to a theoretical mutation parameter. In this section we will discuss the strengths and weaknesses of each method and compare predictions using these methods to predictions using MCMC.

Fitting by Least Squares Regression to Measured Binding Energy Values

To obtain a value for α using least squares regression, we first define a least-squares function $f(\alpha)$ as

$$f(\alpha) = \sum_{i=1}^{n} \left(\Delta\varepsilon_{meas,i} - \alpha\Delta\varepsilon_{pred,i} - \Delta\varepsilon_{wt}\right)^2, \tag{4.11}$$

where $\Delta\varepsilon_{meas}$ is the measured binding energy for an operator mutant and $\Delta\varepsilon_{pred}$ is the corresponding binding energy prediction from our unscaled energy matrix. To determine the best-fit value of α, we identify the value of α that minimizes the function. We perform this fit using measurements from the nine single base pair mutants used in this work.

Fitting to the Average Energy per Mutation

In many cases, we will not have thermodynamic models available to use for inferring scaling factors by fitting or Bayesian inference. This raises the question of whether it is possible to estimate the scaling factor by other means, for example by determining some average binding penalty incurred by making a mutation to a binding site. To explore how we might think about such an average binding penalty, we consider the effects of mutations away from the lowest-energy binding sequence for LacI (Figure 4.8). As shown in Figure 4.8A, a wide range of binding energies are available to binding site mutants. The distribution of binding penalties of single base-pair mutations to this binding site is shown in Figure 4.8B. The distribution is fairly broad, yet we find that the mean predicted binding energy for binding site mutants, as shown in Figure 4.8C, is strongly related to the mean binding penalty of a single base pair mutation. Specifically, the slope of the predicted energy versus

the number of mutations is approximately equal to the mean binding penalty of a single mutation. This tells us that the average energy per mutation is a meaningful metric that provides information about the general behavior of a transcription factor binding site.

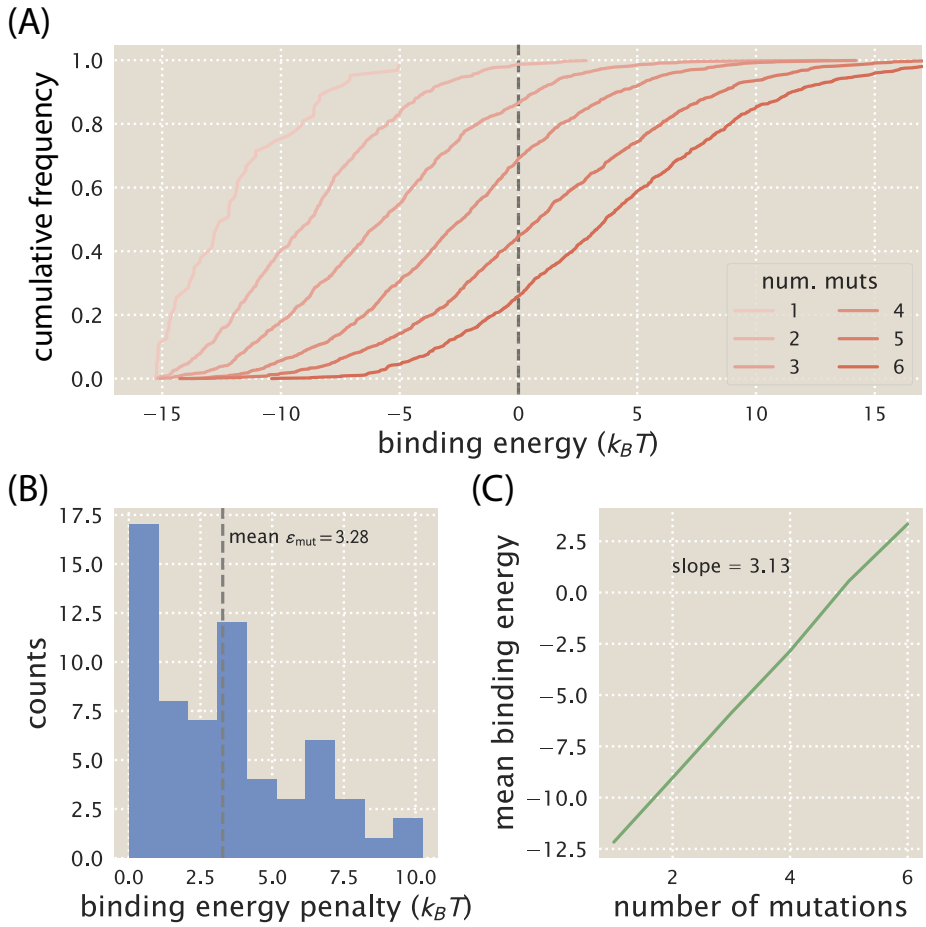

Figure 4.8: **Average effect of a binding site mutation.** (A) Cumulative distributions are shown for the predicted binding energies of *lac* operator mutants. The mean predicted binding energy increases substantially with the number of mutations, as does the width of the distribution. The dotted line shows the point at which $\Delta\varepsilon_R = 0\, k_BT$, which is the average energy of nonspecific binding. (B) A histogram of binding penalties for single base pair mutations to the minimum-energy LacI binding sequence shows that the mean binding penalty of a mutation is $3.28\, k_BT$. (C) Plotting the mean binding energy of an operator against the number of mutations relative to the minimum-energy sequence shows a linear trend with a slope approximately equal to the average energy penalty per mutation.

Next we need to determine how one would estimate the average energy per mutation for an energy matrix that has not already been converted into absolute energy

mutants. We turn to Ref. [61] in which they make an estimate for the average energy penalty, ε_{mut}, of a single base pair mutation relative to the minimum-energy sequence. We can use this estimate to infer a value for α when no thermodynamic model is available to perform a fit for α. We note that unlike the other methods for obtaining α, this method does not rely on expression information from the promoter of interest and thus is best interpreted as a rough "guess."

To begin this estimate, we assume a minimal organism in which there is a single transcription factor with a copy number of 1, and this transcription factor regulates gene expression by binding to a single minimum-energy operator, which has an energy of $\Delta\varepsilon_{min}$. The remaining sequence in this minimal genome is mostly random, but it includes a number of weaker binding sites for the transcription factor such that all possible single base-pair mutations to the binding site are represented. From a statistical mechanics perspective, in order for the transcription factor to bind reliably to the minimum-energy operator, the operator's statistical weight (given by $e^{-\beta\Delta\varepsilon_{min}}$) must outweigh the total statistical weight of all possible single base-pair binding site mutants (given by $le^{-\beta(\Delta\varepsilon_{min}+\varepsilon_{mut})}$), where l is the length of the binding site in base pairs. This gives us

$$e^{-\beta\Delta\varepsilon_{min}} \geq le^{-\beta(\Delta\varepsilon_{min}+\varepsilon_{mut})}. \tag{4.12}$$

This implies that the minimum average binding energy penalty due to a mutation is given by $\varepsilon_{mut} = \ln l$, which for a binding site of 21 bp (the length of a *lac* operator) comes out to $\varepsilon_{mut} \approx 3\ k_BT$. This is remarkably close to the mean energy penalty of $3.28\ k_BT$ calculated for LacI as noted in Figure 4.8B.

Based on this estimate, one can find a value for α by setting the minimum binding energy of an energy matrix to 0, then taking the mean of the nonzero elements of the matrix, ε_{mean}, and finding a scaling factor α such that $\alpha\varepsilon_{mean} = \ln l\ k_BT$.

Method comparison

Each of the methods outlined above is capable of producing a value of α that can be used to convert an energy matrix into k_BT units. Each of these methods has its own advantages and disadvantages. Here we will outline these trade-offs and compare the accuracy of the predictions that can be made using each method.

The primary advantage of the Bayesian regression by MCMC method, which is used for the LacI binding energy predictions in the main text, is that it can be implemented

using the same Sort-Seq data that was used to obtain the energy matrices. No further data collection is required. However, in order to implement this method one must have a thermodynamic model that predicts gene expression for a given operator binding energy. This is trivial for systems with simple regulatory architectures, as is the case with the simple repression architecture used in this study. However, while models for more complex architectures have been proposed [28], identifying the correct model may not be straightforward and a number of additional experiments may be required in order to validate the proposed model. Additionally, significant computing power is required in order to infer a scaling factor using this method.

The advantages of the least-squares fitting method are that it is conceptually straightforward, it requires little computing power, and it provides a very accurate scaling factor. However, multiple fold-change measurements for different operator mutants are required to perform the regression and calculate the best-fit value of α, and any outliers must be identified in order to maximize the accuracy of the fit. Additionally, a thermodynamic model for the system is again required if binding energies are to be measured using fold-change data.

The advantage of the theoretical mutation parameter method is that it is very simple, and requires no knowledge of the regulatory architecture of the promoter. All that is needed is an energy matrix for an operator and an estimate of the operator's length. Indeed, for XylR we lack sufficient information to confidently infer a thermodynamic model of gene expression, so this is the method used to produce energy matrices for this transcription factor (we note that the theoretical mutation parameter method is also used for PurR energy matrices in the main text, though a thermodynamic model is available for PurR as shown in Chapter 3). For the *lac* operator it produces a scaling factor that is approximately as accurate as the other inference methods discussed here (see Figure 4.9). However, this method is based on simplified biophysical arguments, and it is likely that there are a number of regulatory scenarios for which it would not be as successful.

Figure 4.9 compares predictions made using each method for obtaining a scaling factor. The same matrix was used for each prediction, with O1 as the wild-type sequence and $R = 130$ LacI tetramers in the strain used for Sort-Seq. We find that all methods produce predictions that generally describe the data, but when comparing the mean squared error (MSE) of the predictions, it is clear that some methods perform better than others. Note that elsewhere in the supplement we compare predictions using the Pearson correlation coefficient (ρ). We use the

MSE here instead because an inaccurate scaling factor will not effect the linear relationship between predictions and measurements, but it will affect the accuracy of the predictions. Thus a set of predictions may have a high ρ value corresponding to a strong linear relationship, but still have a high MSE corresponding to inaccurate predictions. The Bayesian parameter estimation (Figure 4.9A) and least-squares regression (Figure 4.9B) methods perform nearly identically. However, while the value for α that was inferred from the theoretical mutation parameter (Figure 4.9C) makes predictions that generally describe the data, the MSE values associated with its predictions are notably larger than the other methods, particularly for the 1 bp mutants.

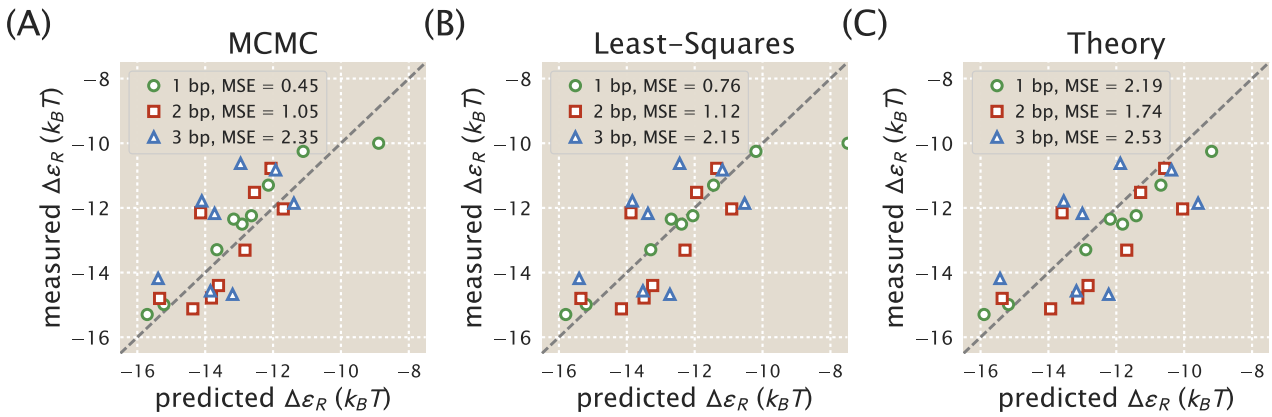

Figure 4.9: **Alternate methods of obtaining energy matrix scaling factor produce similar results.** Shown are data for predicted vs. measured binding energies of 1, 2, or 3 bp mutants. The binding energy predictions are made using energy matrices that have been scaled using one of three methods: (A) Bayesian parameter estimation using MCMC, (B) least-squares regression, or (C) inference from a theoretical mutation parameter. All predictions were made using an energy matrix with O1 as the wild-type sequence and $R = 130$ LacI tetramers in the cells used to perform Sort-Seq. The mean squared error (MSE) associated with each set of predictions is noted in the legend.

4.8 Supplemental Information: Comparing linear energy matrix models with higher-order models

A commonly cited problem with energy matrices is that linear binding models do not accurately describe the mechanism of transcription factor binding to DNA. While linear models assert that each base pair contributes independently to the binding energy, it is known that interactions between two or more base pairs can play an important role in determining binding affinity [32, 62]. In spite of this, linear models are still commonly used because they often perform nearly as well as higher-order models [22, 63], and they require many fewer parameters than a higher-order model. For example, a linear model for LacI binding to a 21 bp long operator requires that 84 parameters be inferred, one for each base at each position. By contrast, a two-point model for LacI that accounts for all possible interactions between any two bases in the binding site requires 3660 parameters. Obtaining high-quality estimates for these parameters requires a great deal more data and computing power than inferring parameters for linear models. Thus it is important to carefully consider whether higher-order models will dramatically improve predictions.

Here we take advantage of our large Sort-Seq data sets to infer two-point binding energy models for LacI binding. As with linear models, two-point binding energy models are inferred by identifying a set of parameters that maximizes mutual information between sequence and expression bin (see Supplemental Section 4.6 for more details). In Figure 4.10 we compare binding energy measurements to a predictions from a linear model (Fig. 4.10(A)) and a two-point model (Fig. 4.10(B)). We also make this comparison for models in which each sequence with only one sequencing count are removed from the data set and then all other sequences are weighted equally 4.10(C-D)). This weighting scheme removes possible sequencing errors from the data set and then gives low-frequency sequences the same influence as high-frequency sequences, compensating for any inequalities that may arise if the library itself has an unequal representation of sequences. The same data set was used to infer each model, namely the data set for the strain with repressor copy number $R = 130$ and an O1 reference sequence. The quality of the predictions for each model is quantified by noting the Pearson's correlation coefficient ρ for each data set. Surprisingly, the unweighted two-point model does not outperform the linear model. In fact, it performs substantially worse. The weighted two-point model, however, performs better than the weighted linear model.

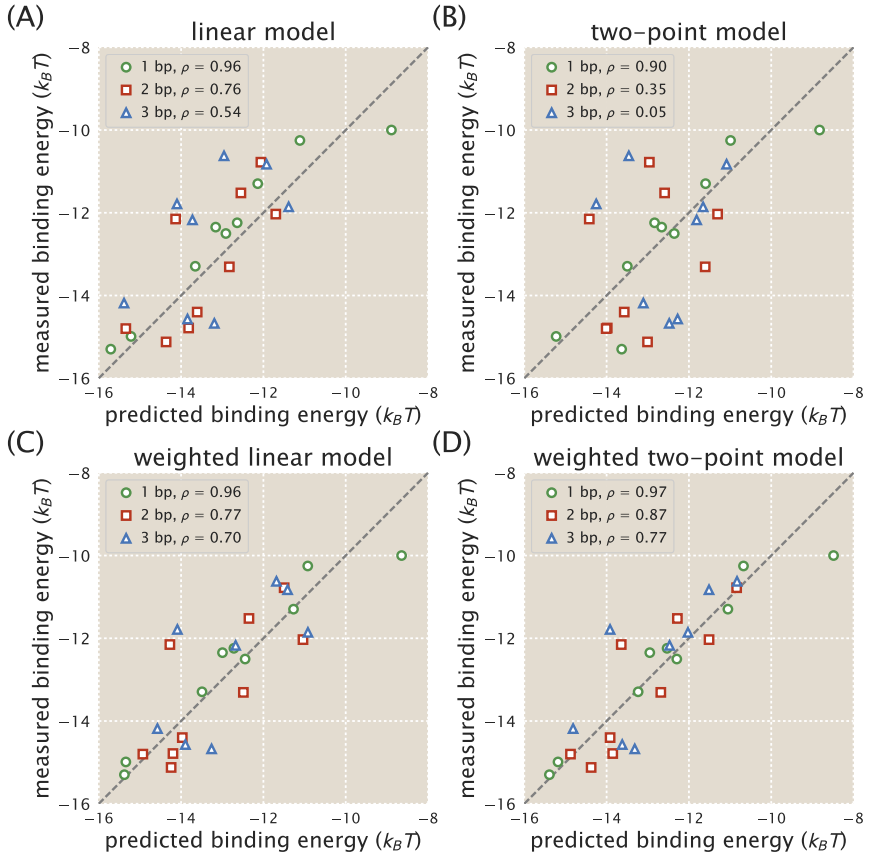

Figure 4.10: **A comparison of linear models with two-point models.** Binding energy measurements are compared against predictions from energy matrix models obtained using a strain where $R = 130$ and O1 is the reference sequence. (A) Predictions are made using a linear energy matrix in which each sequence position is considered independently. This matrix is used to obtain the predictions discussed in the main text. (B) Predictions are made using an energy matrix model that accounts for all two-point interactions between nucleotides at different sequence positions. The Pearson's correlation coefficients for the measurements and predictions indicate that this matrix model performs substantially worse than the linear matrix model, particularly for multiple mutations. (C) Predictions are again made using a linear matrix model, though this model has been weighted so that all sequences (aside from single-count sequences, which were dropped) have the same weight. This matrix model has been inferred after removing all single-count sequences from the data set and then weighting all sequences evenly. (D) Predictions are made using a two-point matrix model using the same weighting scheme as in (C). This weighting procedure results in a two-point matrix model that makes improved predictions relative to the weighted linear matrix model.

4.9 Supplemental Information: Influence of Regulatory Parameters on Energy Matrix Quality

The level of repression in a repressible system is dependent on a number of factors. In this work we primarily focus on operator binding energy, but other key parameters include operator copy number, repressor copy number, and competition from other binding sites, as discussed in detail in Ref. [64]. Here we consider how two parameters influence energy matrix quality: namely, repressor copy number R and the binding energy of the operator reference sequence.

Because our promoter constructs are on plasmids and thus have multiple copies ($N \approx 10$), there is some concern that there might not be a sufficient number of repressors in the cell to demonstrate significant changes in expression when the *lac* operator is mutated. The wild-type copy number of LacI tetramers in *E. coli* is $R = 11$, which is comparable to the plasmid copy number used in this study. We increase the LacI copy number by using synthetic RBSs that have been shown to increase gene expression [29]. Additionally, we consider the fact that the binding energy of the reference sequence influences the distribution of binding energies present in the mutant library, and therefore the "ideal" value of R may be different for different reference sequences. To explore these factors, we performed Sort-Seq experiments for each combination of R (i.e. $R = 30, 62, 130$, or 610) and reference binding energy (i.e. $\Delta\varepsilon_{\mathrm{R}} = -15.3\ k_BT$ for O1, $\Delta\varepsilon_{\mathrm{R}} = -13.9\ k_BT$ for O2, or $\Delta\varepsilon_{\mathrm{R}} = -9.7\ k_BT$ for O3).

Comparison of binding energy predictions

Figure 4.11 shows how predicted and measured binding energy values for single base pair mutants compare for each combination of repressor copy number and reference sequence. We show predictions from energy matrices that have been scaled using the least squares method (see Supplemental Section 4.7), as this is the most accurate method for obtaining a scaling factor. The Pearson's correlation coefficient (ρ) for each set of predictions is shown as a way of quantifying which of these combinations produces the "best" energy matrices, as defined by which matrices give the best agreement between prediction and measurement. We see that the best agreement between prediction and measurement occurs when O1 is the reference sequence. Conversely, predictions from matrices made using O3 as a reference sequence do not predict the measured values at all, as indicated by the especially low ρ values. While the choice of repressor copy number does not appear to have a large effect on the quality of matrix predictions, particularly for matrices with O1 as the reference

sequence, we do observe that $R = 610$ consistently corresponds with the most accurate predictions. We note that in the main text we make predictions using the energy matrix with the O1 reference sequence and $R = 130$. This is because in the main text we obtain our scaling factors using Bayesian inference by MCMC (see Supplemental Section 4.6), and the most accurate scaling factor inferred by this method was for $R = 130$.

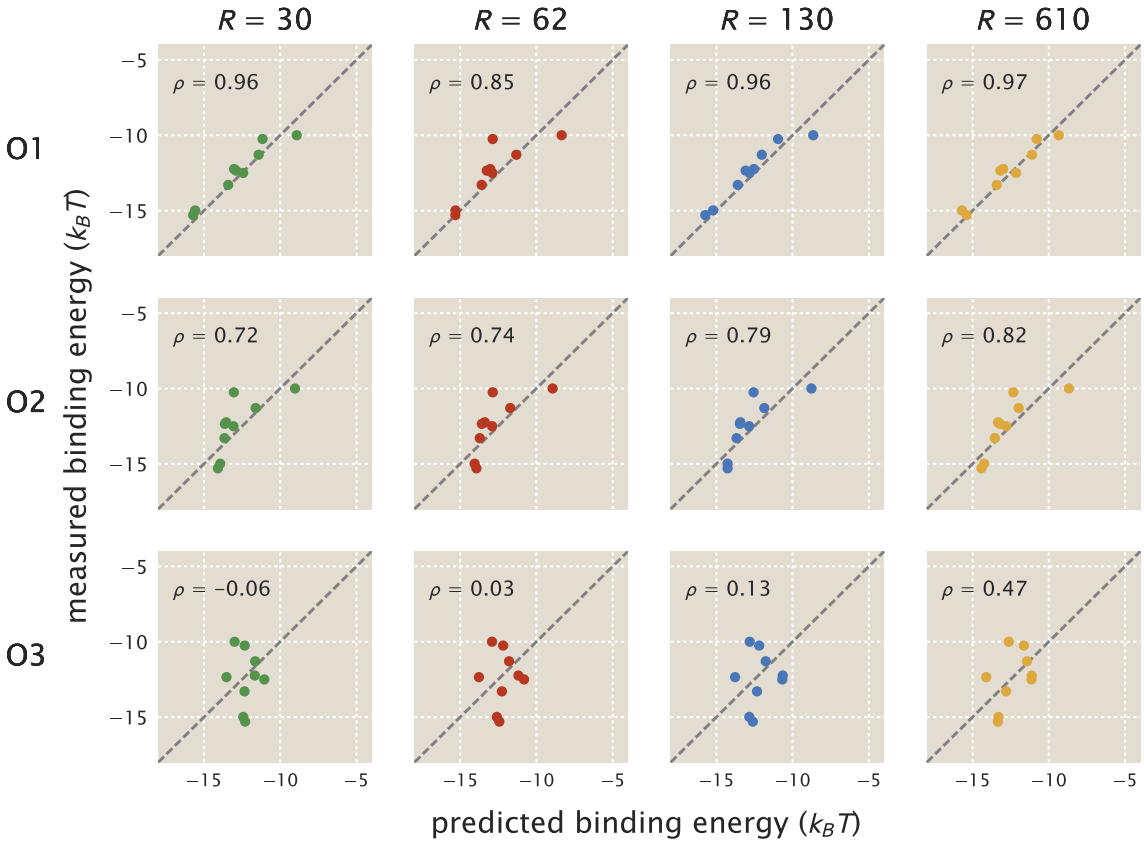

Figure 4.11: **Repressor copy number and reference sequence affect accuracy of energy matrix predictions.** Sort-Seq was performed with all combinations of four different repressor copy numbers ($R = 30, 62, 130$, and 610) and three different reference operator sequences (O1, O2, and O3) to produce a total of 12 energy matrices. Predictions from each of these energy matrices are plotted against measured binding energy values for single base-pair mutants. The Pearson's correlation coefficient (ρ) is noted for each plot as a measure of prediction accuracy.

Variation in energy matrix replicates

In addition to the matrices analyzed in Figure 4.11, we performed two additional replicates for each of the energy matrices obtained from strains with $R = 30$ or $R = 62$. This allows us to determine the level of variation in energy matrices and

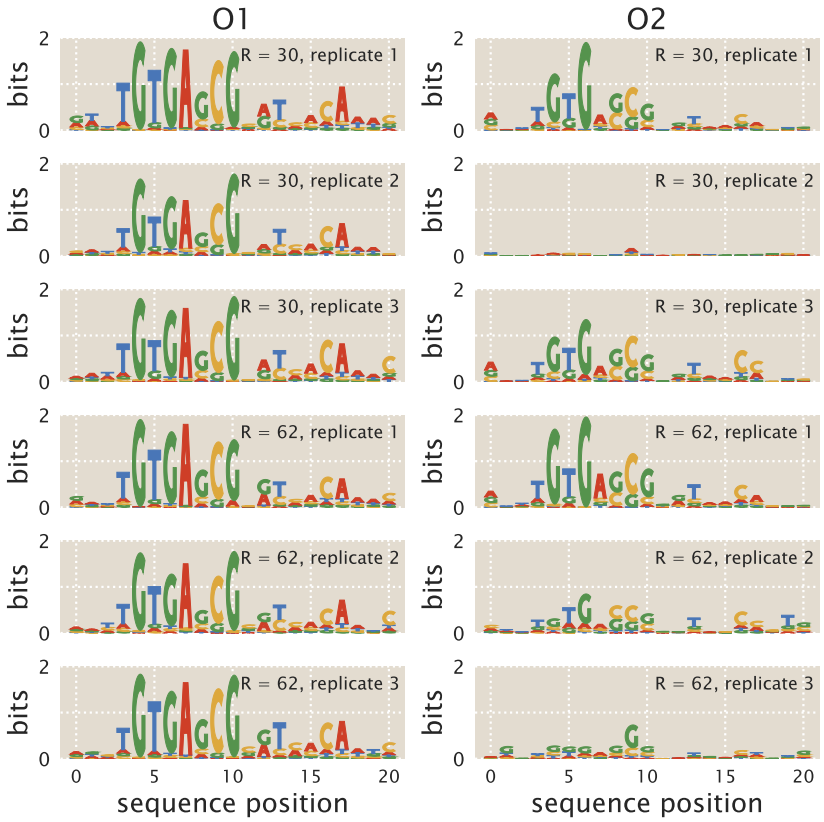

Figure 4.12: **Variation in sequence logo results.** Replicates of Sort-Seq experiments were performed using O1 or O2 as a reference sequence. The O1 experiments (left) produced very consistent sequence logos, while the O2 experiments (right) produced sequence logos that varied significantly in quality.

their associated sequence logos. As shown in Figure 4.12, replicates using O1 as a reference sequence produce very consistent sequence logos, while replicates using O2 as a reference sequence produce less consistent sequence logos. This suggests that the strength of the binding site is a significant factor determining the consistency of experiment outcomes.

As another point of comparison, we computed the Pearson's correlation coefficient ρ between the lists of values comprising each of our unscaled energy matrices with O1 and O2 reference sequences (see Figure 4.13). This allows us to ascertain whether the matrices themselves are substantially different under different experimental conditions. We find that all of the matrices with an O1 reference sequence are highly correlated with one another. By contrast, the matrices with an O2 reference sequence are less correlated with one another, even among replicates of the same experimental conditions. The second replicate of the O2 matrix with $R = 30$

is particularly poorly correlated with other matrices. However, the O2 matrices do generally have a higher ρ value with one another than with the O1 matrices. An exception to this is the O2 matrices with $R = 130$ and $R = 610$, which appear to be moderately well-correlated with the O1 matrices. These results suggest that the choice of reference sequence used to perform the Sort-Seq experiment is a more important determinant of matrix quality than repressor copy number, though the results may also support the hypothesis that higher repressor copy numbers correspond with improved matrix quality, particularly for weaker reference sequences such as O2.

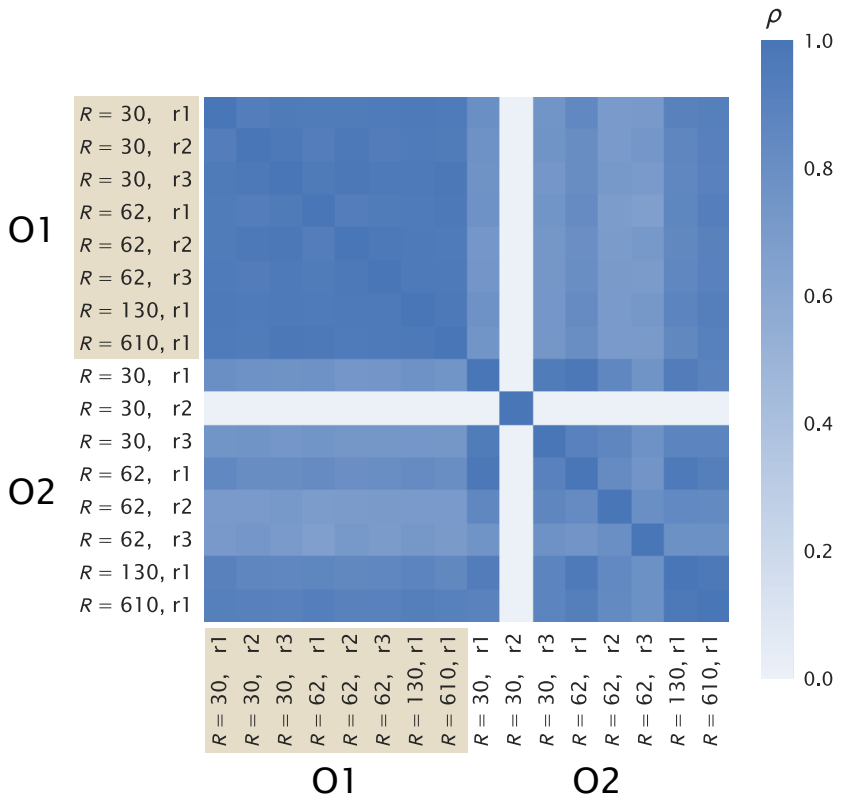

Figure 4.13: **Correlation coefficients between unscaled linear energy matrices.** The Pearson's correlation coefficient (ρ) was calculated for each pair of linear energy matrices with an O1 or O2 reference sequence. Those experiments conducted using strains with repressor copy number $R = 30$ and $R = 62$ were repeated three times, as denoted by replicate number r1, r2, or r3. We find that all O1 matrices are highly correlated with one another, while O2 matrices are generally less correlated with one another. In general there is low correlation between O1 and O2 matrices, with the exception of O2 matrices with high repressor copy numbers, $R = 130$ and $R = 610$.

4.10 Supplemental Information: Comparison of full-promoter and operator-only energy matrix predictions

In the main text, we perform Sort-Seq using libraries in which the entire promoter region was mutated, namely both the RNAP site and the operator. Here we consider whether one can improve energy matrix accuracy by using libraries in which only the operator is mutated.

In order to infer the energy matrix scaling factor α from Sort-Seq data alone (see Supplemental Section 4.6), it is necessary to mutate the full promoter, because mutations to both the operator and RNAP binding sites are relevant to the thermodynamic model used to perform the inference. Because of this we use full promoter mutant libraries in the main text. This means that an alternate method is required in order to infer an energy matrix scaling factor for matrices derived from libraries in which only the operator was mutated. Here, we obtain a scaling factor by least-squares regression to a set of measured binding energies for nine 1 bp mutants, as discussed in Supplemental Section 4.7. We then compare measured binding energies against predictions for 1, 2, and 3 bp mutants that were produced using either full-promoter energy matrices or operator-only energy matrices (see Figure 4.14). All matrices were scaled using a least-squared derived scaling factor. We find that operator-only energy matrices produce somewhat more accurate predictions than full-promoter energy matrices. We quantify this by noting the Pearson correlation coefficient (ρ) for each set of predictions, which clearly indicates that the O1 operator-only matrix produces the most accurate predictions. This shows us that operator-only energy matrices are a good option when it is feasible to infer the scaling factor from binding energy measurements.

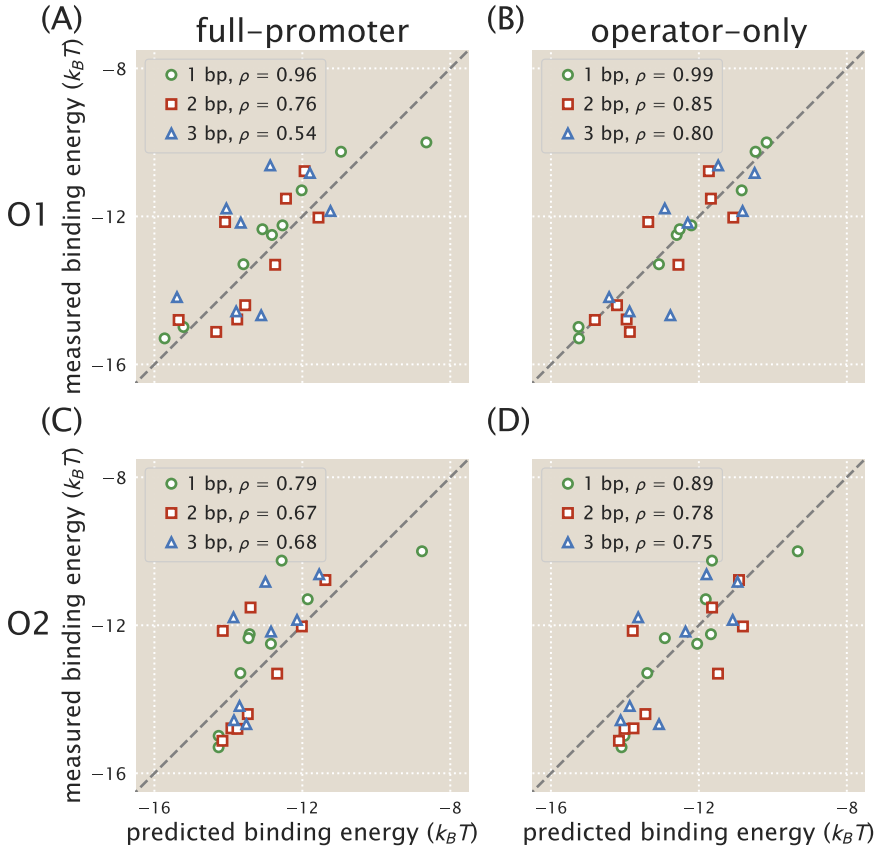

Figure 4.14: **Mutating the operator alone can improve energy matrix accuracy.** Binding energy measurements are plotted against energy matrix predictions from full-promoter (A, C) and operator-only (B, D) energy matrices using either O1 (A, B) or O2 (C, D) as a reference sequence. The Pearson correlation coefficient (ρ) is noted for each set of predictions. We see that the operator-only energy matrices produce more accurate predictions than the full-promoter energy matrices.

4.11 Supplemental Information: Summary of all fold-change data

To measure binding energies for each mutant, fold-change measurements first were obtained by flow cytometry for each mutant in strains with repressor copy numbers $R = 11 \pm 1$, 30 ± 10, 62 ± 15, 130 ± 20, 610 ± 80, and 870 ± 170. The data were fit to the fold-change Equation 4.2. Nonlinear regression was used to obtain the most probable value of $\Delta\varepsilon_R$ for each mutant. The fold-change data, fitted theory curve, and predicted theory curve are shown here for all 1 bp (Figure 4.15), 2 bp (Figure 4.16), and 3 bp mutants (Figure 4.17).

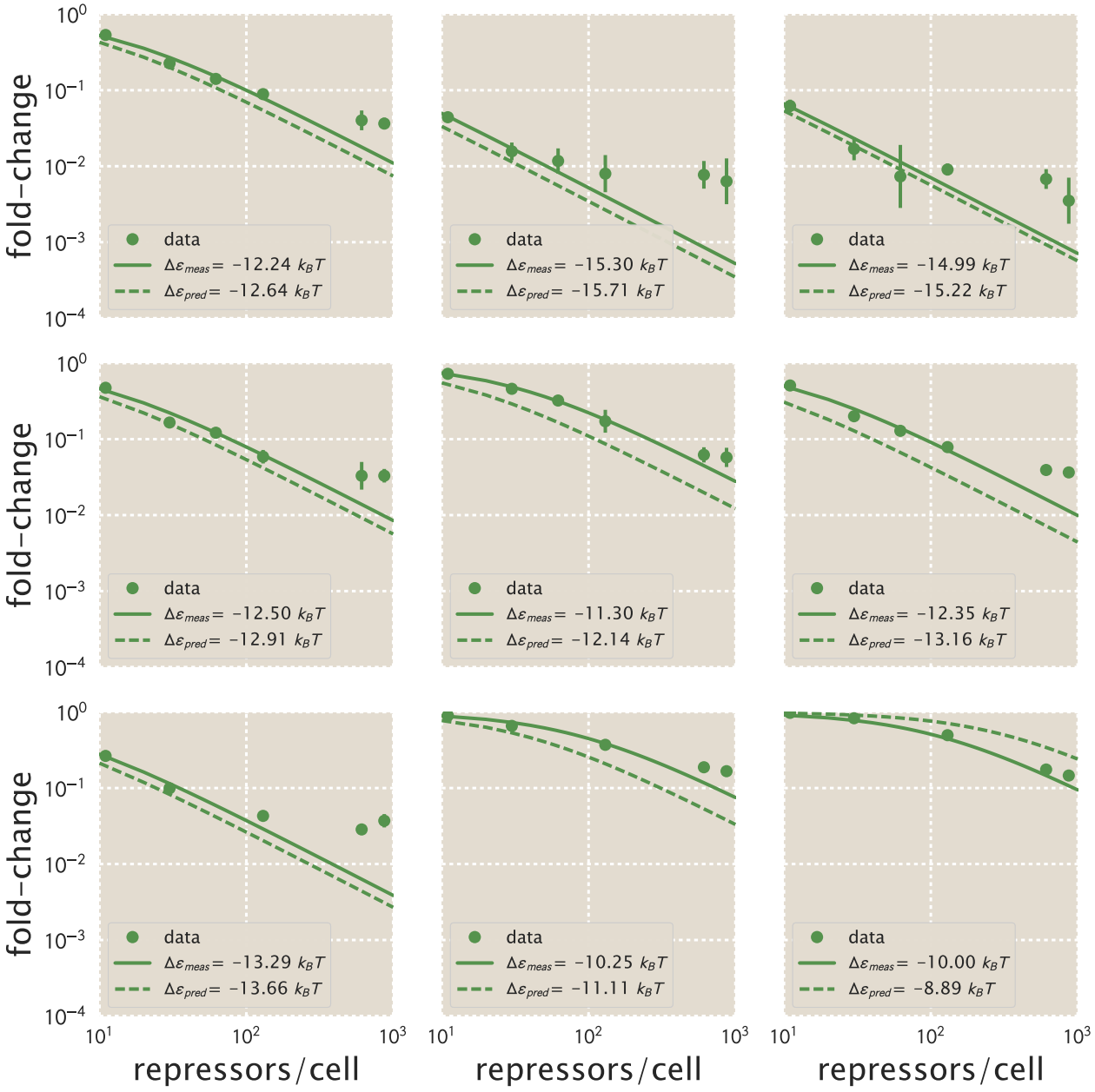

Figure 4.15: **Fold-change measurements for 1 bp mutants.** Fold-change measurements are shown for nine 1 bp operator mutants in strains with $R = 11, 30, 62, 130, 610$, or 870. These measurements are overlaid with the measured (fitted) binding energy measurements for each mutant (solid line) and the predicted measurements (dashed line) as listed in the main text. Note that the bottom three plots do not display data points for $R = 62$, as the data for these strains were outliers.

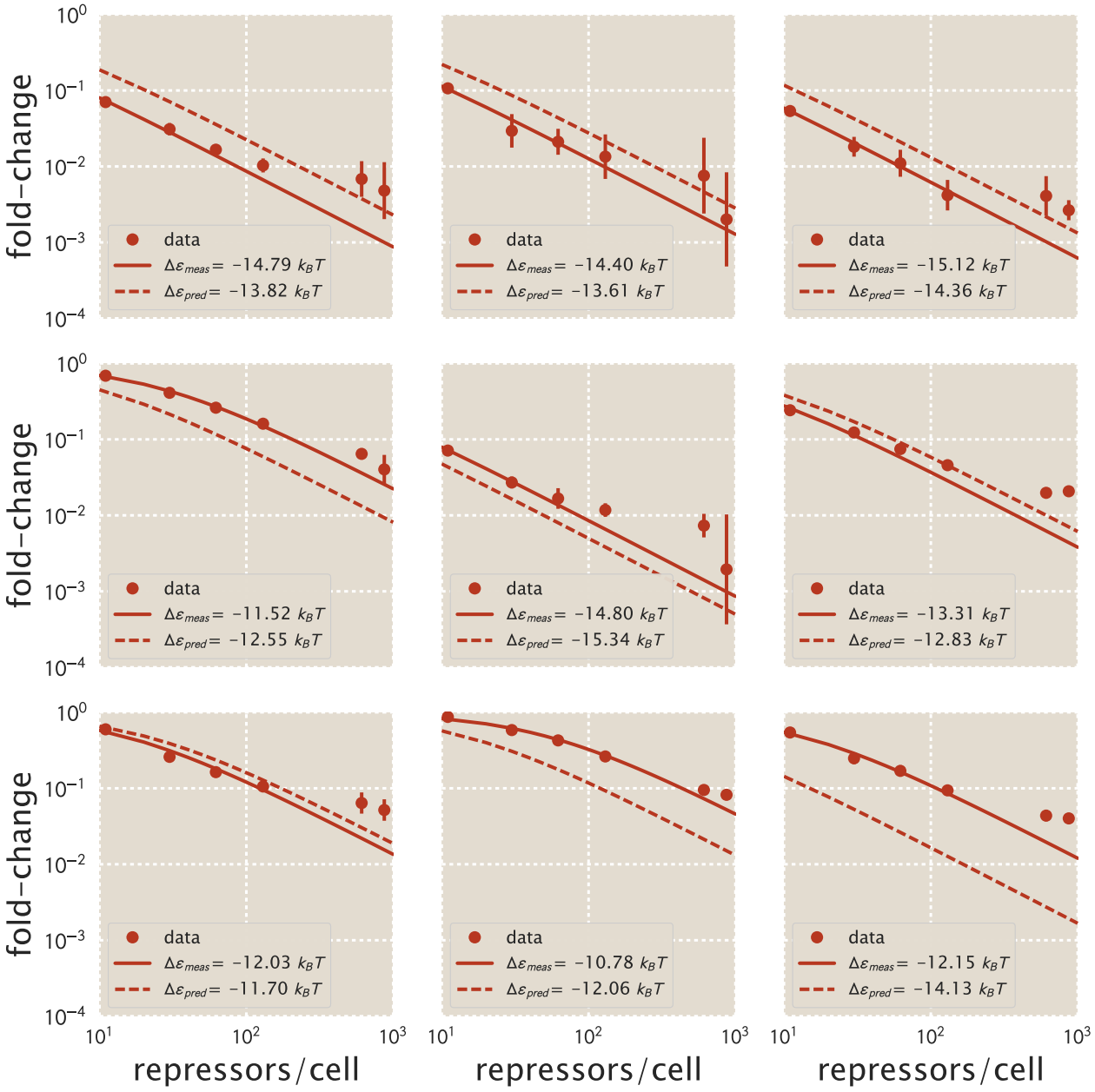

Figure 4.16: **Fold-change measurements for 2 bp mutants.** Fold-change measurements are shown for nine 2 bp operator mutants in strains with $R = 11, 30, 62, 130, 610,$ or 870. These measurements are overlaid with the measured (fitted) binding energy measurements for each mutant (solid line) and the predicted measurements (dashed line) as listed in the main text.

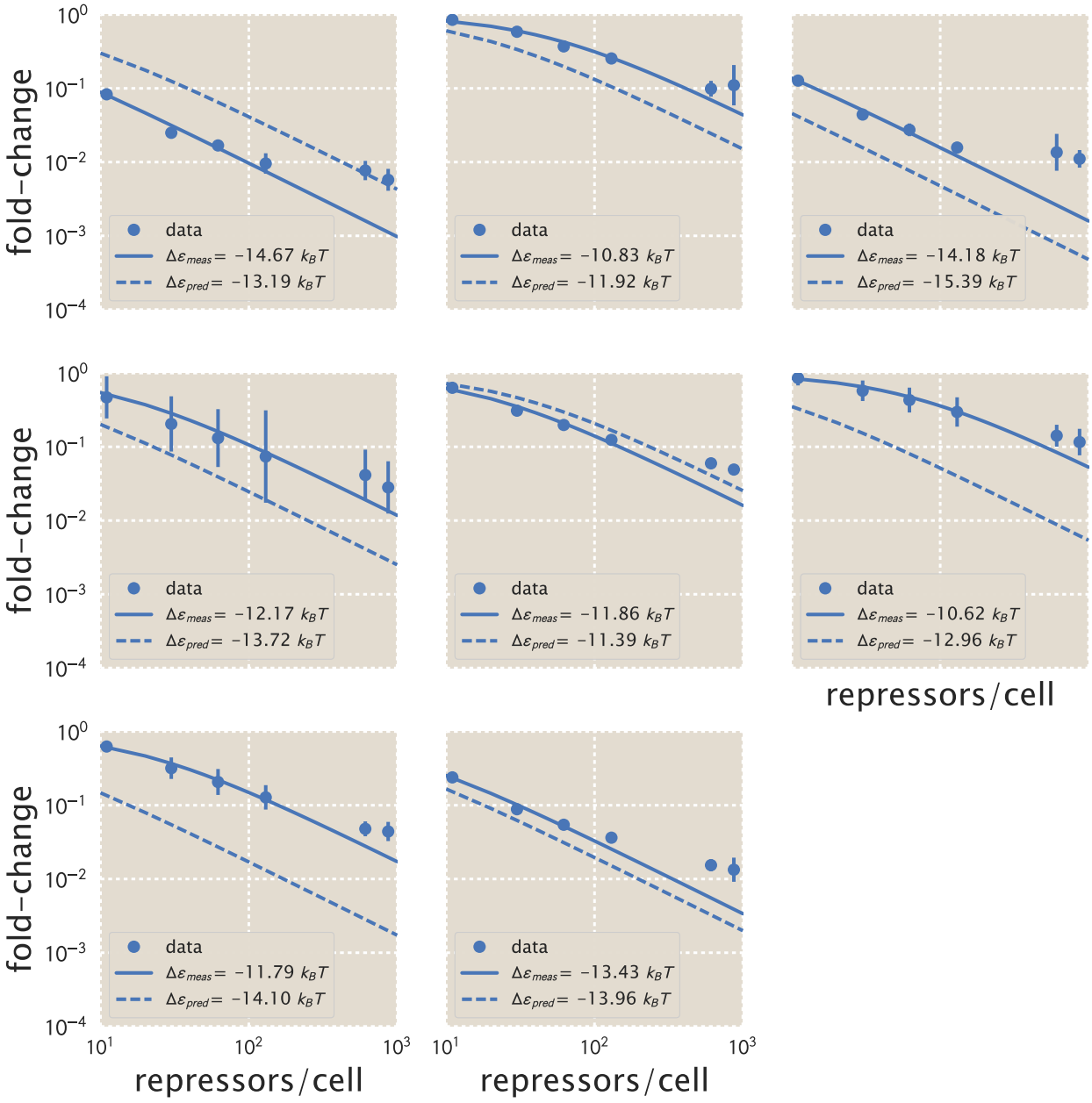

Figure 4.17: **Fold-change measurements for 3 bp mutants.** Fold-change measurements are shown for eight 3 bp operator mutants in strains with $R = 11, 30, 62, 130, 610,$ or 870. These measurements are overlaid with the measured (fitted) binding energy measurements for each mutant (solid line) and the predicted measurements (dashed line) as listed in the main text.

4.12 Supplemental Information: Expressions for phenotypic parameters of induction responses

As discussed in greater detail in Chapter 2, the thermodynamic model we use to predict induction responses allows us to derive expressions for the phenotypic parameters of the induction response. Here we briefly list the expressions for the phenotypic parameters we address in the present work.

The leakiness of the induction curve is the minimum fold-change observed in the absence of ligand, given by

$$\begin{aligned}\text{leakiness} &= (c = 0) \\ &= \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \frac{2R}{N_{NS}} e^{-\beta\Delta\varepsilon_R}\right)^{-1},\end{aligned} \tag{4.13}$$

where c is the concentration of inducer, n is the number of inducer binding sites on the repressor, and $\Delta\varepsilon_{AI}$ is the difference in free energy between the repressor's active and inactive states.

The saturation is the maximum fold change observed in the presence of saturating ligand,

$$\begin{aligned}\text{saturation} &= (c \to \infty) \\ &= \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^n} \frac{2R}{N_{NS}} e^{-\beta\Delta\varepsilon_R}\right)^{-1},\end{aligned} \tag{4.14}$$

where K_A and K_I are the dissociation constants of the inducer and repressor when the repressor is in its active or inactive state, respectively.

Together, these two properties determine the dynamic range of a system's response, which is given by the difference

$$\text{dynamic range} = \text{saturation} - \text{leakiness}. \tag{4.15}$$

The full expression for dynamic range is then given by

$$\text{dynamic range} = \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^n} \frac{2R}{N_{NS}} e^{-\beta\Delta\varepsilon_R}\right)^{-1} - \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \frac{2R}{N_{NS}} e^{-\beta\Delta\varepsilon_R}\right)^{-1}. \tag{4.16}$$

The $[EC_{50}]$ of the induction response denotes the inducer concentration required to generate a system response halfway between its minimum and maximum value such that

$$(c = [EC_{50}]) = \frac{\text{leakiness} + \text{saturation}}{2}. \tag{4.17}$$

The full expression for the $[EC_{50}]$ is then given by

$$\frac{[EC_{50}]}{K_A} = \frac{\frac{K_A}{K_I} - 1}{\frac{K_A}{K_I} - \left(\frac{\left(1+\frac{2R}{N_{NS}}e^{-\beta\Delta\varepsilon_R}\right)+\left(\frac{K_A}{K_I}\right)^n\left(2e^{-\beta\Delta\varepsilon_{AI}}+\left(1+\frac{2R}{N_{NS}}e^{-\beta\Delta\varepsilon_R}\right)\right)}{2\left(1+\frac{2R}{N_{NS}}e^{-\beta\Delta\varepsilon_R}\right)+e^{-\beta\Delta\varepsilon_{AI}}+\left(\frac{K_A}{K_I}\right)^n e^{-\beta\Delta\varepsilon_{AI}}} \right)^{\frac{1}{n}}} - 1. \tag{4.18}$$

BIBLIOGRAPHY

[1] Socorro Gama-castro, Heladia Salgado, Alberto Santos-zavaleta, Daniela Ledezma-tejeida, Luis Mu, Jair Santiago Garc, Kevin Alquicira-hern, Irma Mart, Lucia Pannier, Alejandra Medina-rivera, Hilda Solano-lira, Abraham Castro-mondrag, P. Ernesto, Shirley Alquicira-hern, L. Alejandra, Anastasia Hern, Del Moral-ch, Fabio Rinaldi, and Julio Collado-vides. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 44:133–143, 2016.

[2] S. Oehler, E. R. Eismann, H. Krämer, and B. Müller-Hill. The three operators of the *lac* operon cooperate in repression. *The EMBO journal*, 9(4):973–979, 1990.

[3] Kenn Gerdes, Susanne K. Christensen, and Anders Lobner-Olesen. Prokaryotic toxin-antitoxin stress response loci. *Nature Reviews Microbiology*, 3:371–382, 2005.

[4] Michael N. Alekshun and Stuart B. Levy. Regulation of chromosomally mediated multiple antibiotic resistance: The *mar* regulon. *Antimicrobial Agents and Chemotherapy*, 41(10):2067–2075, 1997.

[5] Stephen D. Minchin and Stephen J.W. Busby. Analysis of mechanisms of activation and repression at bacterial promoters. *Methods*, 47(1):6–12, 2009.

[6] Michal Levo, Tali Avnit-Sagi, Maya Lotan-Pompan, Yael Kalma, Adina Weinberger, Zohar Yakhini, and Eran Segal. Systematic investigation of transcription factor activity in the context of chromatin using massively parallel binding and expression assays. *Molecular Cell*, 65(4):604–617.e6, 2017.

[7] Alexandre Melnikov, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, Andreas Gnirke, Curtis G. Callan, Justin B. Kinney, Manolis Kellis, Eric S. Lander, and Tarjei S. Mikkelsen. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3):271–277, 2012.

[8] Lior Levy, Leon Anavy, Oz Solomon, Roni Cohen, Michal Brunwasser-meirom, Shilo Ohayon, Orn Atar, Sarah Goldberg, Zohara Yakhini, and Roee Amit. A synthetic oligo library and sequencing approach reveals an insulation mechanism encoded within bacterial σ^{54} promoters. *Cell Reports*, 21(3):845–858, 2017.

[9] Michael F. Berger, Anthony A. Philippakis, Aaron M. Qureshi, Fangxue S. He, Preston W. Estep Iii, and Martha L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, 24(11):1429–1435, 2006.

[10] Dana S. Fields, Yi-yuan He, Ahmed Y. Al-Uzri, and Gary D. Stormo. Quantitative specificity of the Mnt repressor. *Journal of Molecular Biology*, 271:178–194, 1997.

[11] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, and Teemu Kivioja. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.

[12] Chaitanya Rastogi, H. Tomas Rube, Judith F. Kribelbauer, Justin Crocker, Ryan E. Loker, Gabriella D. Martini, Oleg Laptenko, William A. Freed-Pastor, Carol Prives, David L. Stern, Richard S. Mann, and Harmen J. Bussemaker. Accurate and sensitive quantification of protein-DNA binding affinity. *Proceedings of the National Academy of Sciences*, In press, 2018.

[13] Sebastian J. Maerkl and Stephen R. Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315:233–238, 2007.

[14] Ryan K. Shultzaberger, Sebastian J. Maerkl, Jack F. Kirsch, and Michael B. Eisen. Probing the informational and regulatory plasticity of a transcription factor DNA-binding domain. *PLoS Genetics*, 8(3), 2012.

[15] Daniel D. Le, Tyler C. Shimko, Arjun K. Aditham, Allison M. Keys, Yaron Orenstein, and Polly Fordyce. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proceedings of the National Academy of Sciences*, In press, 2018.

[16] Cheulhee Jung, John A. Hawkins, Stephen K. Jones, Yibei Xiao, James R. Rybarski, Kaylee E. Dillard, Jeffrey Hussmann, Fatema A. Saifuddin, Cagri A. Savran, Andrew D. Ellington, Ailong Ke, William H. Press, and Ilya J. Finkelstein. Massively parallel biophysical analysis of CRISPR-Cas complexes on next generation sequencing chips. *Cell*, 170(1):35–47, 2017.

[17] Razvan Nutiu, Robin C. Friedman, Shujun Luo, Irina Khrebtukova, David Silva, Robin Li, Lu Zhang, Gary P. Schroth, and Christopher B. Burge. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature Biotechnology*, 29(7):659–664, 2011.

[18] Iris Dror, Tamar Golan, Carmit Levy, Remo Rohs, and Yael Mandel-Gutfreund. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Research*, 25(9):1268–1280, 2015.

[19] Michal Levo, Einat Zalckvar, Eilon Sharon, Ana Carolina Dantas Machado, Yael Kalma, Maya Lotam-Pompan, Adina Weinberger, Zohar Yakhini, Remo Rohs, and Eran Segal. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research*, 25(7):1018–1029, 2015.

[20] Ryan G. Christensen, Ankit Gupta, Zheng Zuo, Lawrence A. Schriefer, Scot A. Wolfe, and Gary D. Stormo. A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. *Nucleic Acids Research*, 39(12):1–9, 2011.

[21] Denise J. Xu and Marcus B. Noyes. Understanding DNA-binding specificity by bacteria hybrid selection. *Briefings in Functional Genomics*, 14(1):3–16, 2015.

[22] Matthew T. Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R. Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, Harmen J. Bussemaker, Morris D. Quaid, Martha L. Bulyk, Gustavo Stolovitzky, Timothy R. Hughes, Phaedra Agius, Aaron Arvey, Philipp Bucher, Curtis G. Callan, Cheng Wei Chang, Chien Yu Chen, Yong Syuan Chen, Yu Wei Chu, Jan Grau, Ivo Grosse, Vidhya Jagannathan, Jens Keilwagen, Szymon M. Kiebasa, Justin B. Kinney, Holger Klein, Miron B. Kursa, Harri Lähdesmäki, Kirsti Laurila, Chengwei Lei, Christina Leslie, Chaim Linhart, Anand Murugan, Alena Myšičková, William Stafford Noble, Matti Nykter, Yaron Orenstein, Stefan Posch, Jianhua Ruan, Witold R. Rudnicki, Christoph D. Schmid, Ron Shamir, Wing Kin Sung, Martin Vingron, and Zhizhuo Zhang. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2):126–134, 2013.

[23] Marko Djordjevic, Anirvan M. Sengupta, and Boris I. Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome Research*, 13:2381–2390, 2003.

[24] Hernan G. Garcia, Alvaro Sanchez, James Q. Boedicker, Melisa Osborne, Jeff Gelles, Jane Kondev, and Rob Phillips. Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. *Cell Reports*, 2(1):150–161, 2012.

[25] Zeba Wunderlich and Leonid A. Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics*, 25(10):434–440, 2009.

[26] Robert C. Brewster, Daniel L. Jones, and Rob Phillips. Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Computational Biology*, 8(12), 2012.

[27] Justin B. Kinney, Anand Murugan, Curtis G. Callan, and Edward C. Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*, 107(20):9158–9163, 2010.

[28] Lacramioara Bintu, Nicolas E. Buchler, Hernan G. Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips. Transcriptional regulation

by the numbers: Models. *Current Opinion in Genetics and Development*, 15(2):116–124, 2005.

[29] Hernan G. Garcia and Rob Phillips. Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences*, 108(29):12173–12178, 2011.

[30] Otto G. Berg and Peter H. Von Hippel. Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology*, 193:723–750, 1987.

[31] Zheng Zuo and Gary D. Stormo. High-resolution specificity from DNA sequencing highlights alternative modes of *lac* repressor binding. *Genetics*, 198(3):1329–1343, 2014.

[32] Matthias Siebert and Johannes Soeding. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*, 44(13):6055–6069, 2016.

[33] Jameson K. Rogers, Christopher D. Guzman, Noah D. Taylor, Srivatsan Raman, Kelley Anderson, and George M. Church. Synthetic biosensors for precise gene control and real-time monitoring of metabolites. *Nucleic Acids Research*, 43(15):7648–7660, 2015.

[34] Julia Rohlhill, Nicholas R. Sandoval, and Eleftherios T. Papoutsakis. Sort-seq approach to engineering a formaldehyde-inducible promoter for dynamically regulated *Escherichia coli* growth on methanol. *ACS Synthetic Biology*, 6(8):1584–1595, 2017.

[35] Tae Seok Moon, Chunbo Lou, Alvin Tamsir, Brynne C. Stanton, and Christopher A. Voigt. Genetic programs constructed from layered logic gates in single cells. *Nature*, 491:249–253, 2012.

[36] James J. Collins, Timothy S. Gardner, and Charles R. Cantor. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, 403(6767):339–342, 2000.

[37] Ertugrul M. Ozbudak, Mukund Thattai, Iren Kurtser, Alan D. Grossman, and Alexander Van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature Genetics*, 31:69–73, 2002.

[38] Nitzan Rosenfeld, Jonathan W. Young, Uri Alon, Peter S. Swain, and Michael B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–1965, 2005.

[39] Priscilla E. M. Purnick and Ron Weiss. The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology*, 10:410–422, 2009.

[40] Leslie Milk, Robert Daber, and Mitchell Lewis. Functional rules for *lac* repressor-operator associations and implications for protein-DNA interactions. *Protein Science*, 19(6):1162–1172, 2010.

[41] Robert Daber, Matthew A. Sochor, and Mitchell Lewis. Thermodynamic analysis of mutant *lac* repressors. *Journal of Molecular Biology*, 409(1):76–87, 2011.

[42] Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of DNA shape in protein–DNA recognition. *Nature*, 461(7268):1248–1253, 2009.

[43] Remo Rohs, Xiangshu Jin, Sean M. West, Rohit Joshi, Barry Honig, and Richard S. Mann. Origins of specificity in protein-DNA recognition. *Annual Reviews in Biochemistry*, 79:233–269, 2010.

[44] Matthew Slattery, Todd Riley, Peng Liu, Namiko Abe, Pilar Gomez-alcala, Iris Dror, Tianyin Zhou, Remo Rohs, Barry Honig, Harmen J. Bussemaker, and Richard S. Mann. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, 147(6):1270–1282, 2011.

[45] Mattias Rydenfelt, Hernan G. Garcia, Robert Sidney Cox, and Rob Phillips. The influence of promoter architectures and regulatory motifs on gene expression in *Escherichia coli*. *PLoS ONE*, 9(12):1–31, 2014.

[46] Jennifer A. N. Brophy and Christopher A. Voigt. Principles of genetic circuit design. *Nature Methods*, 11(5):508–520, 2014.

[47] Ahmad S. Khalil and James J. Collins. Synthetic biology: Applications come of age. *Nature Reviews Genetics*, 11(5):367–379, 2010.

[48] Dominique Bréchemier-baey, Lenin Domínguez-ramírez, and Jacqueline Plumbridge. The linker sequence, joining the DNA-binding domain of the homologous transcription factors, Mlc and NagC, to the rest of the protein, determines the specificity of their DNA target recognition in *Escherichia coli*. *Molecular Microbiology*, 85(5):1007–1019, 2012.

[49] Francisco M. Camas, Eric J. Alm, and Juan F. Poyatos. Local gene regulation details a recognition code within the LacI transcriptional factor family. *PLoS Computational Biology*, 6(11), 2010.

[50] Guillaume Urtecho, Arielle D. Tripp, Kimberly Insigne, Hwangbeom Kim, and Sriram Kosuri. Systematic dissection of sequence elements controlling σ^{70} promoters using a genomically-encoded multiplexed reporter assay in *E. coli*. *Biochemistry*, In press, 2018.

[51] Rupali P. Patwardhan, Choli Lee, Oren Litvin, David L. Young, Dana Pe'Er, and Jay Shendure. High-resolution analysis of DNA regulatory elements by

synthetic saturation mutagenesis. *Nature Biotechnology*, 27(12):1173–1175, 2009.

[52] Eilon Sharon, Yael Kalma, Ayala Sharp, Tali Raveh-Sadka, Michal Levo, Danny Zeevi, Leeat Keren, Zohar Yakhini, Adina Weinberger, and Eran Segal. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, 30(6):521–530, 2012.

[53] Robin P. Smith, Leila Taher, Rupali P. Patwardhan, Mee J. Kim, Fumitaka Inoue, Jay Shendure, Ivan Ovcharenko, and Nadav Ahituv. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics*, 45(9):1021–1028, 2013.

[54] Rolf Lutz and Hermann Bujard. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research*, 25(6):1203–1210, 1997.

[55] Howard M. Salis, Ethan A. Mirsky, and Christopher A. Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, 27(10):946–950, 2009.

[56] Simanti Datta, Nina Costantino, and Donald L. Court. A set of recombineering plasmids for gram-negative bacteria. *Gene*, 379(1-2):109–115, 2006.

[57] Justin Kinney and Gurinder S. Atwal. Parametric inference in the large data limit using maximally informative models. *Neural Computation*, 26(4):637–653, 2014.

[58] Justin B. Kinney, Gašper Tkačik, and Curtis G. Callan. Precise physical models of protein-DNA interaction from high-throughput data. *Proceedings of the National Academy of Sciences*, 104(2):501–506, 2007.

[59] Gurinder S. Atwal and Justin B. Kinney. Learning quantitative sequence–function relationships from massively parallel experiments. *Journal of Statistical Physics*, 162(5):1203–1243, 2016.

[60] William T. Ireland and Justin B. Kinney. MPAthic: quantitative modeling of sequence-function relationships for massively parallel assays. *bioRxiv*, 2016.

[61] Michael Lässig. From biophysics to evolutionary genetics: Statistical aspects of gene regulation. *BMC Bioinformatics*, 8(Suppl 6):S7, 2007.

[62] Panayiotis V. Benos, Martha L. Bulyk, and Gary D. Stormo. Additivity in protein-DNA interactions: How good an approximation is it? *Nucleic Acids Research*, 30(20):4442–51, 2002.

[63] Yue Zhao and Gary D. Stormo. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature Biotechnology*, 29(6):480–483, 2011.

[64] Franz M. Weinert, Robert C. Brewster, Mattias Rydenfelt, Rob Phillips, and Willem K. Kegel. Scaling of gene expression with transcription-factor fugacity. *Physical Review Letters*, 113(25):1–5, 2014.